Survival Analysis with High-Dimensional c \Covariates, with Applications to Cancer Genomics

Dissertation Advisor: Professor Yi Li Sihai Dave Zhao

Survival analysis with high-dimensional covariates, with applications to cancer genomics

Abstract

Recent technological advances have given cancer researchers the ability to gather vast amounts of genetic and genomic data from individual patients. These offer tantalizing possibilities for, for example, basic cancer biology, tailored therapies, and personalized risk predictions. At the same time, they have also introduced many analytical difficulties that cannot be properly addressed with current statistical procedures, because the number of genomic covariates in these datasets is often larger than the sample size. In this dissertation we study methods for addressing this so-called high-dimensional issue when genomic data are used to analyze time-to-event outcomes, so common to clinical cancer studies.

In Chapter 1, we propose a regularization method for sparse estimation for estimating equations. Our method can be used even when the number of covariates exceeds the number of samples, and can be implemented using well-studied algorithms from the non-linear constrained optimization literature. Furthermore, for certain estimating equations and certain regularizers, including the lasso and group lasso, we prove a finite-sample probability bound on the accuracy of our estimator.

However, it is well-known that these types of regularization methods can achieve better performance if a quick and simple procedure is first used to reduce the number of covariates. In Chapter 2, we propose and theoretically justify a principled method for reducing dimensionality in the analysis of censored data by selecting only the important covariates. Our procedure involves a tuning parameter that has a simple interpretation as the desired false

positive rate of this selection.

Similar types of model-based screening methods have also been proposed, but only for a few specific models. Model-free screening methods have also recently been studied, but can have lower power to detect important covariates. In Chapter 3 we propose a screening procedure that can be used with any model that can be fit using estimating equations, and provide unified results on its finite-sample screening performance. We thus generalize many recently proposed model-based and model-free screening procedures. We also propose an iterative version of our method and show that it is closely related to a recently studied boosting method for estimating equations.

Contents

Acknowledgments

Thanks to my advisor, Yi Li, for his prescient direction, his generous time, and his unwavering commitment to my success.

Thanks to my committee members Giovanni Parmigiani and Nikhil Munshi, for challenging me to think sharply and rigorously, for teaching me about successful collaboration, and for including me in their academic families.

Thanks to the administrative staff for tolerating, and obliging, my frequent requests for help.

And thanks to my professors, classmates, friends, and family.

Sparse estimation for estimating equations using decomposable norm-based regularizers

Sihai Dave Zhao and Yi Li

Department of Biostatistics
Harvard School of Public Health

1.1 Introduction

With the advent of high-dimensional datasets, sparse estimation has become increasingly important in regression modeling. However, traditional methods such as stepwise selection have been found to be unstable (Breiman, 1996), resulting in poor predictive performance. Recent interest has focused on regularization methods such as the lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), or group lasso (Yuan and Lin, 2006), which can perform simultaneous estimation and variable selection. These procedures have also proved useful in analyzing the increasingly common high-dimensional data setting, where the number of covariates p can be larger than the number of subjects n.

Most such procedures are based on regularizing some objective function, such as a likelihood. However, in many cases full likelihood models are difficult to specify, especially with more complicated data structures such as correlated observations or censored data. In other situations, such as for robust estimation, it is desirable to model only the first or the first few moments of the data instead of the entire likelihood.

For example, in Section 1.3 we study short-term survival in patients with early-stage non-small-cell lung cancer using SNP data and other clinical covariates. We are interested in selecting factors that might be associated with the probability of surviving fewer than 3 years, so that we can preemptively identify higher-risk patients and treat them with more aggressive therapies. However, we have data on only 100 patients but 74,666 SNPs. Furthermore, we cannot apply the lasso or SCAD because the model we use to predict 3-year survival does not fit under the likelihood framework due to our censored outcomes.

In Section 1.3 we also study a multicenter clinical trial of radiation therapy against two schedules of concurrent chemoradiotherapy for head and neck cancer. Patients on the standard radiation therapy arm on average do worse than patients on one of the chemoradiotherapy arms, but also experience significantly fewer toxicities. We are interested in selecting

Table 1.1: A taxonomy of regularization algorithms

	Regularizing function	
Estimation method	ℓ_1-*type*	*More complex*
Objective function	Coordinate descent LARS	Coordinate descent (convex) MM algorithms (nonconvex)
Estimating equations	Penalized estimating equations Smooth-thresholding EEBoost Penalized quadratic form	Penalized estimating equations Penalized quadratic form

factors that might be associated with better survival within patients on the standard arm. Patients predicted to survive longer then do not need to be treated with the more toxic therapy. Traditional methods for subgroup analysis often encounter difficulties due to multiple testing (Lagakos, 2006), and these can be alleviated by using a multivariate variable selection procedure such as the lasso. In this case, several of the factors are categorical and need to be grouped together when selecting variables, but we cannot use the group lasso because patient survival times are correlated between centers, making a likelihood difficult to specify.

In the absence of a likelihood, estimation is often carried out with estimating equations. There has been a great deal of interest in regularization procedures for estimating equations that can achieve sparse estimation (see Table 1.1), but some difficulties remain. Fu (2003), Johnson et al. (2008), and Wang et al. (2011) studied penalized estimating equations, but their proposals can be computationally demanding, or may not yield exactly sparse estimates (Zhang et al., 2010). Ueki (2009) developed an ℓ_1-norm-type regularization methods based on smooth thresholding, but requires initial $n^{1/2}$-consistent parameter estimates, which are difficult to obtain with high-dimensional covariates. Most recently, Wolfson (2011) developed a boosting algorithm (EEBoost) for estimating equations and showed that this procedure can produce solutions similar to those from ℓ_1-penalized methods, but EEBoost cannot accommodate more complex regularizing functions, such as the group lasso penalty function. Finally, none of these papers consider finite-sample properties of their estimators.

A popular alternative approach to regularization in the absence of a likelihood is to penalize an objective function so that the problem is shifted back into the first row of Table 1.1. For example, Wang et al. (2008) regularize the semiparametric AFT model by penalizing the Buckley-James least-squares loss function. For the semiparametric linear transformation model, Zhang et al. (2010) proposed constructing a quadratic form from the estimating equation, using the inverse of an estimate of the estimating equation covariance matrix. They implemented this penalized quadratic form (PQF) method by using an iterative procedure, where at each step they penalized a pseudo-least-squares problem derived from a linear approximation to their quadratic form. In this fashion the PQF can use existing algorithms to regularize estimating equations using both simple and complex regularizing functions. However, the covariance matrix can be difficult to estimate and invert, especially if p is large.

In this paper we propose the nonlinear constrained optimization selector (NCOS), a new sparse estimation procedure for estimating equations. The method is appealing in that:

1. It requires no initial estimates and can be used with a wide variety of both simple and complex regularizing functions.

2. For the appropriate regularizing functions, it can perform simultaneous estimation and variable selection.

3. When the true parameter vector is sparse and the regularizing function is a *decomposable norm*, a concept introduced by Negahban et al. (2009) and reviewed in Section 1.5.2, we can give a finite-sample bound on the ℓ_2-error of its estimates for a certain class of estimating equations.

4. It can be implemented using a variety of well-studied algorithms from the nonlinear constrained optimization literature.

When the estimating equation is the score equation of the linear model, and when the

regularizing function is the ℓ_1-norm, our NCOS will reduce to the Dantzig selector (Candès and Tao, 2007). In previous work, James and Radchenko (2009), Antoniadis et al. (2010), and Johnson et al. (2011) studied the Dantzig selector for generalized linear models, the Cox model, and the AFT model, respectively, but offered no theory applicable to other types of estimating equations. Dicker (2011) recently extended the Dantzig selector to generalized estimating equations and proved asymptotic results, but required a consistent initial estimator of the regression coefficients. Finally, all of these methods used the ℓ_1-norm as a regularizer. Liu et al. (2009) proposed and theoretically justified a version of the Dantzig selector with a group-based regularizer, but their work is limited to the ordinary linear model.

We introduce our NCOS and describe its implementation in Section 1.2, and we apply it to the non-small-cell lung cancer and head and neck cancer studies in Section 1.3. In Section 1.4 we study its performance in simulations, and we give our theoretical error bound in Section 1.5. We conclude with a few remarks in Section 1.6, and leave all proofs for the Appendix.

1.2 Nonlinear constrained optimization selector

1.2.1 Estimating equations

Let $Y_i = (Y_{i1}, \ldots, Y_{iK_i})^T$ be a $K_i \times 1$ outcome vector and $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iK_i})^T$ be a $K_i \times p$ matrix of covariates for units $i = 1, \ldots, n$, where the K_i are independent and identically distributed discrete random variables. When $K_i = 1$ for all i, our formulation reduces to the usual independent data setting.

Let the $p \times 1$ vector $\boldsymbol{\beta}_0$ be the true parameter vector, and let $\mathbf{U}(\boldsymbol{\beta})$ be an estimating function from $\mathbb{R}^p$ to $\mathbb{R}^p$ that depends on $(Y_i, \mathbf{X}_i), i = 1, \ldots, n$ such that $\mathrm{E}\{\mathbf{U}(\boldsymbol{\beta}_0)\} = 0$. When $p < n$, $\boldsymbol{\beta}_0$ is usually estimated by finding the $\hat{\boldsymbol{\beta}}$ such that $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$. When $p > n$,

however, there are an infinite number of $\boldsymbol{\beta}$ that can solve $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$, introducing the need for a regularizing function $r(\boldsymbol{\beta})$ to choose between the different possible solutions.

1.2.2 Method

We may consider estimating $\boldsymbol{\beta}_0$ by choosing the $\boldsymbol{\beta}$ with the smallest $r(\boldsymbol{\beta})$ that also satisfies $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$. However, requiring $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$ typically leads to overfitting. We instead consider "almost solving" the estimating equation, or requiring that $\|\mathbf{U}(\boldsymbol{\beta})\|_\infty \leq \gamma$ for some $\gamma > 0$.

More specifically, our NCOS estimate is defined as the $\hat{\boldsymbol{\beta}}$ that solves

$$\text{minimize } r(\boldsymbol{\beta}) \text{ subject to } \|\mathbf{U}(\boldsymbol{\beta})\|_\infty \leq \gamma. \tag{1.1}$$

For example, if $\gamma = \|\mathbf{U}(\mathbf{0})\|_\infty$, then $\boldsymbol{\beta} = \mathbf{0}$ is a feasible estimate, and if $r(\boldsymbol{\beta})$ reaches a minimum at $\mathbf{0}$, $\hat{\boldsymbol{\beta}}$ will be shrunk to $\mathbf{0}$. The shrinkage parameter γ trades off between the bias and variance of $\hat{\boldsymbol{\beta}}$, and can be chosen using some tuning procedure such as generalized cross-validation or cross-validation. Note that (1.1) reduces to the Dantzig selector when $r(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ and $\mathbf{U}(\boldsymbol{\beta})$ is the least-squares score equation.

1.2.3 Implementation

Because $r(\boldsymbol{\beta})$ and $\mathbf{U}(\boldsymbol{\beta})$ can be nonlinear functions, we implement the NCOS using methods from nonlinear optimization (Leyffer and Mahajan, 2010), which gives our proposal its name. For computational readiness we here employ a sequential linear programming strategy, an iterative procedure which at each step solves a linear programming subproblem using approximations to the nonlinear objective and constraint functions. This iterative approach must be combined with a mechanism for ensuring global convergence, so we implement the trust region and filter method of Huang et al. (2011).

In particular, we first define $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}^+, \boldsymbol{\beta}^-)$, where $\beta_j^+ = \max(0, \beta_j)$ and $\beta_j^- = \max(0, -\beta_j)$ such that each component of $\tilde{\boldsymbol{\beta}}$ is positive. We then define the functions $\tilde{r}(\tilde{\boldsymbol{\beta}}) = r(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-)$ and $\tilde{\mathbf{U}}(\tilde{\boldsymbol{\beta}}) = \mathbf{U}(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-)$. Finally, we begin with an initial value $\tilde{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$ and at each iteration k solve the subproblem

$$\text{minimize } \tilde{r}(\tilde{\boldsymbol{\beta}}^{(k)}) + \nabla\tilde{r}(\tilde{\boldsymbol{\beta}}^{(k)})^T\mathbf{d} \tag{1.2}$$

$$\text{subject to } \|\tilde{\mathbf{U}}(\tilde{\boldsymbol{\beta}}^{(k)}) - \tilde{\mathbf{I}}(\tilde{\boldsymbol{\beta}}^{(k)})^T\mathbf{d}\|_\infty \leq \gamma, \|\mathbf{d}\|_1 \leq \tau^{(k)}, \tilde{\beta}_j + d_j > 0, j = 1, \ldots, 2p, \tag{1.3}$$

where we optimize over $\mathbf{d}$. Here $\tilde{\mathbf{I}}(\tilde{\boldsymbol{\beta}}) = -\partial\tilde{\mathbf{U}}/\partial\tilde{\boldsymbol{\beta}}$ and $\tau^{(k)}$ is the radius of the "trust region" in which the linear approximations are deemed appropriate. The solution $\mathbf{d}^{(k)}$ gives us a trial point $\tilde{\boldsymbol{\beta}}^{(k)} + \mathbf{d}^{(k)}$. If this trial point meets certain conditions, it is accepted and $\tilde{\boldsymbol{\beta}}^{(k+1)}$ is updated. Otherwise, $\tilde{\boldsymbol{\beta}}^{(k+1)} = \tilde{\boldsymbol{\beta}}^{(k)}$. At the end of each iteration, the size of the trust region is readjusted; see (Huang et al., 2011), who also prove that this method is globally convergent. When $\tilde{r}(\tilde{\boldsymbol{\beta}})$ is not continuously differentiable at $\mathbf{0}$, we replace $\nabla\tilde{r}(\mathbf{0})$ by a subgradient at $\mathbf{0}$. Our estimate $\hat{\boldsymbol{\beta}}$ will equal $\boldsymbol{\beta}^{+(k)} - \boldsymbol{\beta}^{-(k)}$ at convergence.

If the initial value $\tilde{\boldsymbol{\beta}}^{(0)}$ is too far from the true minimizer, the constraints of the linear subproblem can be incompatible. To address this we follow the warm starts strategy of Friedman et al. (2007) and compute NCOS estimates for decreasing values of the tuning parameter, using the solution for one value of γ as the initial estimate for the next. We let $\gamma_{\max} = \|\mathbf{U}(\mathbf{0})\|_\infty$ and decrease γ on the logarithmic scale to $\gamma_{\min} = \epsilon\gamma_{\max}$, where ϵ is small, for example 10^{-4}.

1.3 Data examples

1.3.1 Non-small-cell lung cancer

Huang et al. (2009) conducted a genome-wide SNP analysis of tumor cells from subjects with early-stage non-small-cell lung cancer to find predictors for overall survival. The study

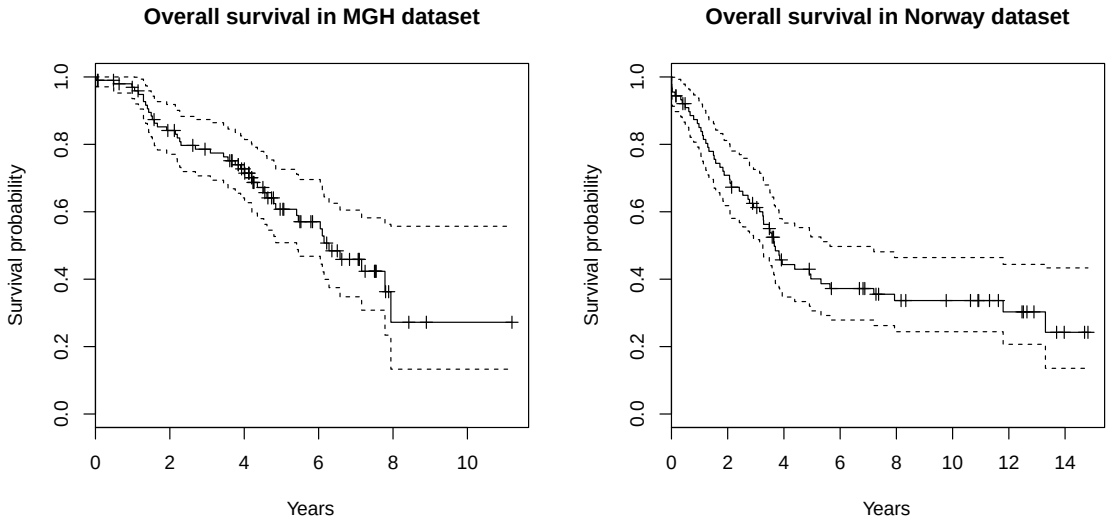

Figure 1.1: Overall survival in MGH and Norway datasets

population consisted of $n = 100$ patients who underwent surgical resection at Massachusetts General Hospital (MGH). The Kaplan-Meier estimate of the survival curve is displayed in Figure 1.1. After selecting 74,666 SNPs with $\geq 95\%$ call rate, $\geq 10\%$ subjects with heterozygous or variant homozygous alleles, and $\geq 3\%$ subjects with variant homozygous for analysis, they found five SNPs whose prognostic significance they were able to replicate in 89 similar patients in a validation dataset from the Norwegian National Institute of Occupational Health (STAMI), whose Kapalan-Meier estimate is also shown in Figure 1.1. They also measured age, gender, cell type (squamous cell carcinoma or adenocarcinoma), and smoking history (pack-years), and it is of interest to determine which SNPs or clinical covariates are associated with surviving less than 3 years, which is close to the 20^{th} percentile of survival times of the MGH patients. These patients could then be placed on more aggressive or experimental therapies.

To model this 3-year survival we followed Jung (1996). Let T_i be the survival time, C_i be the censoring time, $X_i = \min(T_i, C_i)$, and $\delta_i = I(T_i \leq C_i)$ for the i^{th} patient. The number of risk alleles for the j^{th} SNP is given by Z_{ij}, where risk alleles are defined as those associated

with shorter survival (Huang et al., 2009). We modeled

$$\text{logit}\{\text{P}(T_i \geq t_0 \mid \mathbf{Z}_i)\} = \boldsymbol{\beta}_0^T \mathbf{Z}_i \tag{1.4}$$

for t_0 equal to 3 years. Let $\pi(\eta) = \text{logit}^{-1}(\eta)$ and $\pi'(\eta) = \partial\pi/\partial\eta$. To fit the model, Jung (1996) proposed modifying the logistic regression score equations to handle censored data, under the assumption that the C_i are independent of T_i and $\mathbf{Z}_i$:

$$\mathbf{U}(\boldsymbol{\beta}) = n^{-1} \sum_i^n \frac{\mathbf{Z}_i \pi'(\boldsymbol{\beta}^T \mathbf{Z}_i)}{\pi(\boldsymbol{\beta}^T \mathbf{Z}_i)\{1 - \pi(\boldsymbol{\beta}^T \mathbf{Z}_i)\}} \left\{ \frac{I(X_i \geq t_0)}{\hat{S}_C(t_0)} - \pi(\boldsymbol{\beta}^T \mathbf{Z}_i) \right\}, \tag{1.5}$$

where $\hat{S}_C(t)$ is the Kaplan-Meier estimate of the survival function of the censoring time.

With many more covariates than subjects, we followed the recommendation of Fan and Lv (2008) to preprocess high-dimensional data by screening out a large number of unimportant covariates before using regularization procedures. This common practice helps to improve the speed, stability, and accuracy of the regularization. We therefore first applied their sure independence screening method to reduce the number of SNPs to 500. We also included age, gender, cell type, and smoking history for a total of 504 covariates.

Clearly, (1.5) does not correspond to any likelihood function, but simultaneous estimation and variable selection can be accomplished with a regularization method for estimating equations using an ℓ_1-norm regularizing function. Given the disadvantages of the penalized estimating equation and smooth-thresholding procedures discussed in Section 1.1, Table 1.1 suggests that we use the NCOS, the PQF, or EEBoost. We investigated the performances of all three procedures. Note that the PQF as originally proposed by Zhang et al. (2010) requires inverting the $p \times p$ estimating equation covariance matrix, but this is difficult for large p. In our implementation we replaced this covariance matrix by the identity matrix, such that our PQF penalized the squared ℓ_2-norm of $\mathbf{U}(\boldsymbol{\beta})$.

The tuning parameters for these three methods cannot be chosen using AIC or BIC in the absence of a likelihood. Minimizing the cross-validation estimate of out-of-sample prediction error is another popular option but can be computationally intensive. Here we followed

Table 1.2: Covariates associated with 3-year survival in non-small-cell lung cancer (NCOS)

Covariate/SNP	Gene
Age (years)	
rs2072778	TIPRL: TIP41, TOR signaling pathway regulator-like (S. cerevisiae)
rs13117571	No gene information
rs10739959	No gene information
rs1541871	LSAMP: limbic system-associated membrane protein
rs1514607	No gene information
rs1557689	LHFPL3: lipoma HMGIC fusion partner-like 3
rs11190065	CNNM1: cyclin M1
rs9829162	PDZRN3: PDZ domain containing ring finger 3
rs6549543	PDZRN3: PDZ domain containing ring finger 3
rs16884956	No gene information
rs2422705	No gene information
rs2826217	No gene information
rs7832451	No gene information
rs1769792	No gene information
rs17493316	CAMK1D: calcium/calmodulin-dependent protein kinase 1D
rs6989777	NRG1: neuregulin 1
rs17050678	CCRN4L: CCR4 carbon catabolite repression 4-like (S. cerevisiae)
rs978927	ADAMTS3: ADAM metallopeptidase with thrombospondin type 1 motif, 3
rs13219662	No gene information
rs4941229	No gene information
rs6751438	No gene information
rs12822507	CREBL2: cAMP responsive element binding protein-like 2
rs7973428	GPR19: G protein-coupled receptor 19

Johnson et al. (2008) and minimized the simple generalized cross validation-type criterion $\widehat{BS}/(1-n^{-1}\|\hat{\boldsymbol{\beta}}\|_0)^2$, where $\widehat{BS}$ is an estimate of the in-sample Brier score at t_0. If $\hat{\pi}(t_0 \mid \mathbf{Z}_i)$ is the survival probability at t_0, conditional on $\mathbf{Z}_i$, of patient i predicted by our fitted model, the Brier score is defined in (Graf et al., 1999) as

$$BS = n^{-1}\sum_i \left[\frac{\{0-\hat{\pi}(t_0 \mid \mathbf{Z}_i)\}^2 N_i(t_0)}{\hat{S}_C(X_i)} + \frac{\{1-\hat{\pi}(t_0 \mid \mathbf{Z}_i)\}^2 Y_i(t_0)}{\hat{S}_C(t_0)}\right]. \tag{1.6}$$

We used $\|\hat{\boldsymbol{\beta}}\|_0$, the number of non-zero coefficients of $\hat{\boldsymbol{\beta}}$, to estimate the degrees of freedom of the fitted model.

We compared the NCOS, the PQF, and EEBoost by using the models fitted on the MGH data to produce risk scores for patients in the STAMI dataset, which we then used to

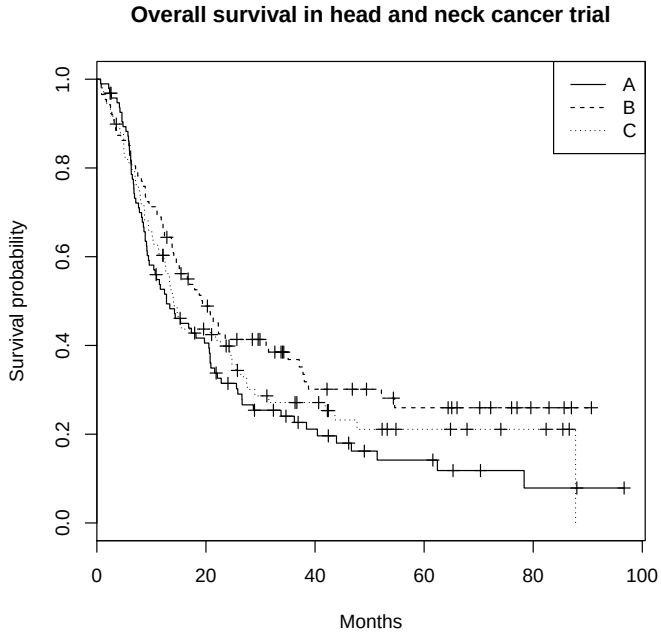

Figure 1.2: Overall survival in head and neck cancer trial; arm A: radiation therapy, arms B and C: two schedules of concurrent chemoradiotherapy

calculate the AUC statistic for surviving less than three years (Uno et al., 2007). The NCOS, the PQF, and EEBoost gave AUC statistics of 0.573, 0.482, and 0.577, respectively. Though the NCOS and EEBoost performed similarly with respect to prediction, the NCOS produced the smallest model, selecting 24 covariates. In contrast, the PQF selected 40 covariates and EEBoost selected 35. Table 1.2 lists the covariates found by NCOS to be predictive of short-term survival, and interestingly some proteins associated with these SNPs, such as the ADAM metallopeptidases, are thought to be involved in the molecular biology of non-small-cell lung cancer (Zhou et al., 2006).

1.3.2 Head and neck cancer

Adelstein et al. (2003) conducted a multicenter randomized clinical trial of radiation therapy alone (arm A) against two schedules of concurrent chemoradiotherapy (arms B and C) for patients with nonresectable head and neck cancer. The Kaplan-Meier estimates of the overall

survival curves are given in Figure 1.2. They found that patients on arm A performed significantly worse than patients on arm B, but that the latter group experienced significantly more toxicities. They also collected various patient characteristics, and it is of interest to determine which can be used to predict the patients on arm A who are likely to see good results, so that they do not need to receive the more toxic treatment. We considered age, height, weight, sex, race (white, black or other), ECOG performance status (Oken et al., 1982), primary tumor site (oral cavity, oropharynx, hypopharnyx, or larynx), tumor differentiation (well-differentiated, moderately well-differentiated, or poorly or undifferentiated), tumor extent (T1 to T3 or T4), nodal status (N0 to N2 or N3), smoking history (pack-years), and alcohol consumption (ounces per week).

We modeled these data using a marginal Cox model with a working independence correlation structure, following Spiekerman and Lin (1998). The patients on arm A came from $n = 50$ different institutions, with between 1 and 30 patients per instution for a total of 271 subjects. Let T_{ik} be the survival time, C_{ik} be the censoring time, $X_{ik} = \min(T_{ik}, C_{ik})$, and $\delta_{ik} = I(T_{ij} \leq C_{ik})$ for the k^{th} patient at the i^{th} institution. Also define $N_{ik}(t) = I(X_{ik} \leq t, \delta_{ik} = 1)$ and $Y_{ik}(t) = I(X_{ik} \geq t)$. We modeled the marginal hazard function for T_{ik} as $\lambda(x; \mathbf{Z}_{ik}) = \lambda_0(x) \exp(\boldsymbol{\beta}_0^T \mathbf{Z}_{ik})$, where the baseline hazard function is shared across institutions, and we used the estimating equation

$$\mathbf{U}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K_i} \int \left\{ \mathbf{Z}_{ik} - \frac{\sum_{i=1}^{n} \sum_{k=1}^{K_i} \mathbf{Z}_{ik} Y_{ik}(x) \exp(\boldsymbol{\beta}^T \mathbf{Z}_{ik})}{\sum_{i=1}^{n} \sum_{k=1}^{K_i} Y_{ik}(x) \exp(\boldsymbol{\beta}^T \mathbf{Z}_{ik})} \right\} dN_{ik}(x). \tag{1.7}$$

Our semiparametric model and correlated outcomes made it difficult to specify a likelihood or partial likelihood for (1.7). Furthermore, many of our covariates are categorical, so we needed to use the group lasso penalty function to effect variable selection. Table 1.1 suggests that we use the NCOS or the PQF. We investigated the performance of both procedures, where again in our PQF we penalized the squared ℓ_2-norm of $\mathbf{U}(\boldsymbol{\beta})$. We tuned the NCOS and the PQF to minimize the C-statistic (Uno et al., 2011a) calculated from five-fold cross-validation.

Table 1.3: Covariates associated with overall survival in head and neck cancer

Covariate	NCOS	PQF
Age (years)	0	0
Height (cm)	0.02	0
Weight (kg)	0	0
Sex (male)	-0.48	0
Race (white)	0	0
Race (black)	0	0
Performance status	0.09	0
Tumor site (oral cavity)	0	0
Tumor site (oropharynx)	0	0
Tumor site (hypopharynx)	0	0
Tumor differentiation (well-differentiated)	0.41	0
Tumor differentiation (moderately well-differentiated)	-0.06	0
T4	0	0
N3	0	0
Smoking (pack-years)	0.12	0.07
Alcohol (ounces per week)	0.06	0

We note that in previous work Cai et al. (2005) developed variable selection methodology specifically for the marginal multivariate Cox model under the penalized objective function framework. They applied the SCAD penalty (Fan and Li, 2001) to a pseudo-partial likelihood and derived asymptotic properties. However, their penalization approach cannot be extended to arbitrary estimating equations.

To evaluate the predictive ability of the regularization methods, we used five-fold cross validation of the data from arm A to estimate the out-of-sample C-statistics of the estimated models. In each fold we used cross-validation in the training set for tuning. Our NCOS gave an average C-statistic of 0.64, with a standard deviation of 0.03, while the PQF gave an average C-statistic of 0.54, with a standard deviation of 0.05. Thus the model produced by the NCOS provides a better risk classification for patients on arm A. We applied both methods to all patients on arm A, and Table 1.3 gives the resulting parameter estimates. Note that our NCOS selected the indicator variables corresponding to race, tumor site, and tumor differentiation in an all-in or all-out fashion.

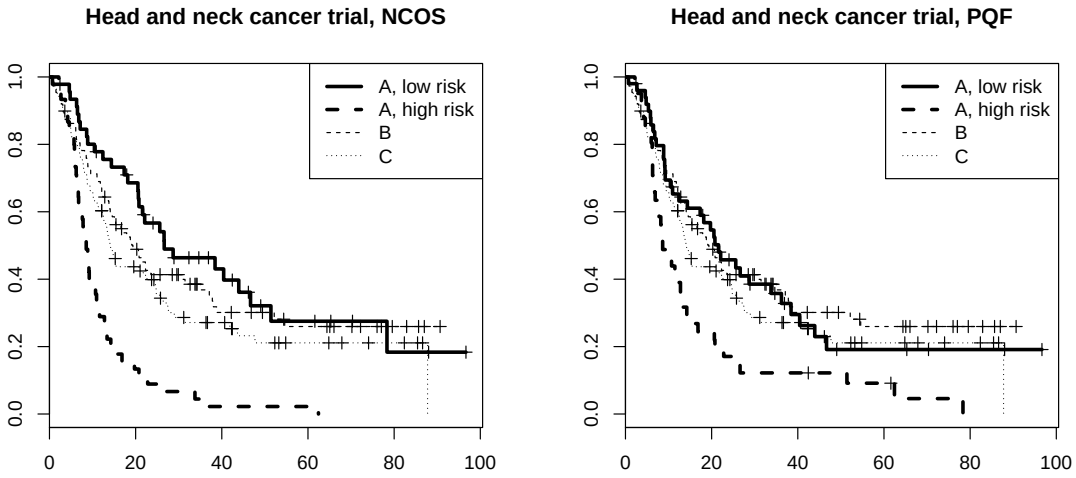

Figure 1.3: Risk classifications using the NCOS and PQF parameter estimates

As an illustration of the risk discrimination abilities of the models fit using the NCOS and the PQF, we used the coefficients in Table 1.3 to calculate risk scores for each patient in arm A. We then dichotomized the patients into those above and below the median risk scores, and plotted Kaplan-Meier estimates of the resulting risk groups in Figure 1.3. Though these plots do not represent the out-of-sample predictive ability of the two methods, we can see that at least in the training data, the NCOS model identifies a subgroup of patients on arm A who perform better than even the patients on arm B.

1.4 Simulation results

1.4.1 Short-term survival model with ℓ_1-norm

In these simulations we studied the regularization of the short-term survival estimating equation (1.5) with an ℓ_1-norm regularizer, as described in Section 1.3.1. We simulated $n = 200$ subjects, and for each subject we generated either $p = 10$ or $p = 500$ covariates, to mimic low- and high-dimensional cases, from a zero-mean multivariate normal

Table 1.4: Results for the short-term survival model, $p = 10$

Method	ρ	Size	FN	FP	MSE	AUC
NCOS	0.5	6.8 (2)	0.15 (0.19)	0.57 (0.26)	0.89 (0.67)	0.75 (0.05)
PQF	0.5	7.11 (2.01)	0.15 (0.19)	0.62 (0.25)	0.91 (0.61)	0.75 (0.05)
EEBoost	0.5	7.14 (1.85)	0.14 (0.17)	0.62 (0.25)	0.9 (0.66)	0.75 (0.05)
NCOS	0.9	4.87 (2.43)	0.41 (0.27)	0.42 (0.29)	2.8 (3.5)	0.66 (0.07)
PQF	0.9	4.42 (2.41)	0.43 (0.26)	0.35 (0.29)	1.99 (1.75)	0.67 (0.07)
EEBoost	0.9	5.47 (2.28)	0.34 (0.25)	0.47 (0.29)	2.9 (3.77)	0.66 (0.06)

with covariate matrix $\sigma_{ij} = \rho^{|i-j|}$, where ρ equaled either 0.5 or 0.9. When $p = 10$ we let $\boldsymbol{\beta}_0 = (-1.1, 0.5, 0, 0, -0.3, 0, 0, 0, 0.8, 0)^T$, and when $p = 500$ we let $(\beta_{01}, \ldots, \beta_{010})$, $(\beta_{031}, \ldots, \beta_{040})$, and $(\beta_{061}, \ldots, \beta_{070})$ equal the low-dimensional $\boldsymbol{\beta}_0$, with $\beta_{0j} = 0$ for all other j.

For both the low- and high-dimensional cases we simulated survival data from $\log(T_i) = \boldsymbol{\beta}_0^T \mathbf{Z}_i + \varepsilon_i$ with ε_i having a logistic distribution with mean -0.5 and scale 1. Under this scheme the model of Jung (1996) is correctly specified. Finally, we generated C_i from an independent exponential distribution to give approximately 50% censoring.

We observed that the 20^{th} percentiles of the simulated survival times were roughly $t_0 = 0.1$ when $p = 10$ and $t_0 = 0.05$ when $p = 500$, and we used these times when implementing (1.5). We simulated 200 datasets and calculated the average sizes of the estimated models, the average false negative (FN) and false positive (FP) rates, and empirical mean squared errors (MSEs). Finally, to evaluate out-of-sample performance we calculated AUC statistics, using our fitted models to predict the probability of surviving past t_0 in independent test datasets. We studied the NCOS, the PQF, and EEBoost using an ℓ_1 regularizer. To tune the methods, we used the generalized cross validation-type criterion from Section 1.3.1.

The results are reported in Tables 1.4 and 1.5. In the low-dimensional case, the three methods appear very similar, though the NCOS selects on average smaller models while still achieving good MSEs and AUCs. In the high-dimensional case, however, the NCOS significantly outperforms the PQF and EEBoost. When $\rho = 0.5$, the NCOS selects two-

Table 1.5: Results for the short-term survival model, $p = 500$

Method	ρ	Size	FN	FP	MSE	AUC
NCOS	0.5	63.72 (13.58)	0.5 (0.12)	0.12 (0.03)	5.6 (1)	0.67 (0.06)
PQF	0.5	91.9 (7.66)	0.48 (0.12)	0.18 (0.02)	8.08 (1.42)	0.63 (0.06)
EEBoost	0.5	81.3 (11.4)	0.48 (0.12)	0.15 (0.02)	57.12 (13.86)	0.65 (0.07)
NCOS	0.9	28.16 (16.96)	0.81 (0.14)	0.05 (0.03)	6.64 (0.72)	0.59 (0.07)
PQF	0.9	44.23 (26.57)	0.79 (0.16)	0.09 (0.05)	7.43 (1.07)	0.58 (0.07)
EEBoost	0.9	74.66 (11.54)	0.66 (0.13)	0.14 (0.02)	77.22 (18.84)	0.59 (0.07)

thirds as many covariates as the other methods, yet does not have a significantly higher FN rate. Furthermore, its performance in MSE and AUC is comparable to that of the PQF and better than that of EEBoost. When $\rho = 0.9$, EEBoost selects large models with high MSEs, while the NCOS and the PQF look more similar. However, the NCOS now outperforms the PQF in every category. These results are perhaps because for the ℓ_1-norm regularizer, the sequential linear programming algorithm we use for the NCOS is guaranteed to be globally convergent (Huang et al., 2011), which is not the case for the PQF. The large MSEs of EEBoost in high dimensions are surprising, though the performance of EEBoost in this realm has not yet been well-explored.

1.4.2 Multivariate Cox model with group lasso

In these simulations we studied the regularization of the multivariate Cox model (1.7) with the group lasso, as described in Section 1.3.2. We simulated data from $n = 50$ clusters and let the number of subjects per cluster K_i have a discrete uniform distribution between 1 and 4. For the i^{th} subject in the k^{th} cluster we generated latent variables V_{ikj}, with $j = 1, \ldots, 15$ or $j = 1, \ldots, 100$ to mimic low- and high-dimensional cases. The V_{ikj} came from a zero-mean multivariate normal with covariance matrix $\sigma_{ij} = \rho^{|i-j|}$, where ρ equaled either 0.5 or 0.9. Following Yuan and Lin (2006), we trichotomized each V_{ikj} as 0, 2, or 1 if it was less than $\Phi^{-1}(1/3)$, greater than $\Phi^{-1}(2/3)$, or in between, respectively, to give covariate vectors $\mathbf{Z}_{ik} = \{I(V_{ik1} = 0), I(V_{ik1} = 1), I(V_{ik2} = 0), I(V_{ik2} = 1), \ldots\}^T$. In the low-dimensional case, where $p = 30$, we let $\boldsymbol{\beta}_0$ be all zero except for $(\beta_{01}, \beta_{02}, \beta_{05}, \beta_{06}, \beta_{09}, \beta_{010})^T =$

Table 1.6: Results for the multivariate Cox model, $p = 30$

Method	ρ	Size	FN	FP	MSE	C-statistic
NCOS	0.5	12.42 (10.94)	0.27 (0.3)	0.35 (0.4)	5.7 (4.53)	0.74 (0.03)
PQF	0.5	14.86 (10.98)	0.3 (0.31)	0.44 (0.39)	6.36 (6.3)	0.74 (0.03)
NCOS	0.9	11.78 (9.9)	0.26 (0.28)	0.33 (0.36)	6.04 (8.31)	0.74 (0.04)
PQF	0.9	13.58 (8.09)	0.19 (0.26)	0.36 (0.3)	5.4 (6.3)	0.75 (0.04)

$(1.2, -1.8, -0.5, -1, -1, -1)^T$. In high dimensions, where $p = 200$, we let $(\beta_{01}, \ldots, \beta_{030})$ and $(\beta_{031}, \ldots, \beta_{060})$ equal the low-dimensional $\boldsymbol{\beta}_0$, with $\beta_{0j} = 0$ for all other j.

We simulated correlated survival data from $\log(T_{ik}) = \boldsymbol{\beta}_0^T \mathbf{Z}_{ik} + \varepsilon_{ik}$, and we let $\varepsilon_{ik} = F^{-1}\{\Phi^{-1}(N_{ik})\}$, where $F(x) = 1 - \exp(-e^x)$ is the CDF of the standard Gumbel distribution, $\Phi(\cdot)$ is the standard normal CDF, and $(N_{i1}, \ldots, N_{iK_i})$ comes from a multivariate normal with mean zero and covariance matrix $\sigma_{ij} = 0.5^{|i-j|}$. Finally, we generated C_{ik} from an independent exponential distribution to give approximately 50% censoring.

We applied the NCOS and the PQF to (1.7), and we used the group lasso regularizer, where the j^{th} group consisted of the variables $I(V_{ij} = 0)$ and $I(V_{ij} = 1)$. We simulated 200 datasets and calculated the average sizes, FN and FP rates, and MSEs. To evaluate out-of-sample performance we calculated the average C-statistics (Uno et al., 2011a) of the fitted models in independently simulated test datasets.

To tune the methods, we used the generalized cross validation-type criterion $\widehat{IBS}/(1 - n^{-1}\|\hat{\boldsymbol{\beta}}\|_0)^2$ for computational convenience. Here IBS is the integrated Brier score of Graf et al. (1999). If $\hat{\pi}(t \mid \mathbf{Z}_i)$ is the survival probability at time t, conditional on $\mathbf{Z}_i$, of patient i predicted by the fitted Cox model, then the IBS is defined as

$$IBS = \int^{t^*} n^{-1} \sum_i \left[\frac{\{0 - \hat{\pi}(t \mid \mathbf{Z}_i)\}^2 N_i(t)}{\hat{S}_C(X_i)} + \frac{\{1 - \hat{\pi}(t \mid \mathbf{Z}_i)\}^2 Y_i(t)}{\hat{S}_C(t)} \right] \frac{dt}{t^*}, \tag{1.8}$$

where $t^* = \max(X_i)$ is the largest observed failure time.

The results are reported in Tables 1.6 and 1.7. The methods show the same trends as in

Table 1.7: Results for the multivariate Cox model, $p = 200$

Method	ρ	Size	FN	FP	MSE	C-statistic
NCOS	0.5	8.06 (4.78)	0.58 (0.15)	0.02 (0.02)	10.93 (1.68)	0.77 (0.06)
PQF	0.5	10.42 (8.52)	0.72 (0.18)	0.04 (0.04)	12.9 (1.97)	0.7 (0.11)
NCOS	0.9	9.94 (6.22)	0.51 (0.17)	0.03 (0.03)	10.54 (1.34)	0.79 (0.04)
PQF	0.9	16.27 (13.61)	0.64 (0.28)	0.06 (0.06)	12.66 (2.66)	0.69 (0.12)

Section 1.4.1. In almost every case, the NCOS selects smaller models. In the low-dimensional case the methods look comparable, though the NCOS tends to have slightly higher FN rates and lower FP rates, which matches our results from Section 1.3.2. In high-dimensions the NCOS again performs better in nearly every category, especially for $\rho = 0.9$. In terms of variable selection, its FN rates are much lower than those of the PQF, and it also has lower FP rates. Most strikingly, it has lower MSEs and higher C-statistics, with smaller variability in both statistics.

1.5 Theoretical error bound

We have mentioned that the Dantzig selector can be viewed as a special case of our NCOS. One appealing feature of the Dantzig selector is that finite-sample probability bounds on the ℓ_2-error of the estimator can be calculated, and naturally the question arises as to whether the same can be done for our NCOS.

Unfortunately, we have found it difficult to derive probability bounds for any arbitrary estimating equation and any arbitrary regularizing function. First, most estimating equations are nonlinear, so the methods used for deriving the bound on the Dantzig selector do not immediately apply. Second, bounds of this type require conditions like the restricted isometry property (RIP) of Candès and Tao (2007), which is used to limit the collinearity of the model. However, in general the collinearity of an estimating equation-based regression model depends not only on the covariates but also on $\boldsymbol{\beta}$, so that any RIP-type conditions

must depend on $\boldsymbol{\beta}$ as well. Finally, the special properties of the ℓ_1 penalty function that are necessary for stating these bounds are not necessarily shared by arbitrary $r(\boldsymbol{\beta})$.

However, we will be able to give an exact finite-sample probability bound on the size of $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ for a certain class of estimating equations and a certain class of $r(\boldsymbol{\beta})$. This class of estimating equations includes some important cases, and the class of $r(\boldsymbol{\beta})$ includes the commonly used lasso and group lasso penalties.

1.5.1 Assumptions on $\mathbf{U}(\boldsymbol{\beta})$

Assumption 1 *The estimating equation $\mathbf{U}(\boldsymbol{\beta}_0)$ can be written as $n^{-1}\sum_{i=1}^{n}\boldsymbol{\psi}_i(\boldsymbol{\beta}_0; \mathbf{Y}_i, \mathbf{X}_i)$ for $p \times 1$ vector-valued functions $\boldsymbol{\psi}_i$ that depend on $\mathbf{Y}_i$ and $\mathbf{X}_i$.*

Assumption 2 *The $\boldsymbol{\psi}_i(\boldsymbol{\beta}_0; \mathbf{Y}_i, \mathbf{X}_i)$ have mean $\mathbf{0}$, and for the j^{th} component $\psi_{ij}(\boldsymbol{\beta}_0; \mathbf{Y}_i, \mathbf{X}_i)$ there exist constants M and v such that $\mathrm{E}|\psi_{ij}(\boldsymbol{\beta}_0; \mathbf{Y}_i, \mathbf{X}_i)|^m \leq m!M^{m-2}v/2$ for $m \geq 2$ and for all j.*

Assumption 1 states that $\mathbf{U}(\boldsymbol{\beta})$ is a sum of independent and identically distributed terms. Assumption 2 bounds the moments of those terms, and can usually be satisfied by assuming bounded covariates. While Assumption 1 is somewhat restrictive, it holds for a number of important estimating equations, such as generalized estimating equations (Liang and Zeger, 1986). These assumptions allow us to invoke Bernstein's inequality, but for certain $\mathbf{U}(\boldsymbol{\beta})$ they can be relaxed. For example, Bernstein-type inequalities exist for U-statistics (Hoeffding, 1963) and martingales (van de Geer, 1995).

We must also assume that $\mathbf{U}(\boldsymbol{\beta})$ satisfies something similar to the restricted isometry property (RIP) of Candès and Tao (2007). We will call a vector $\mathbf{c}$ is k-sparse if at most k of its components are nonzero. An arbitrary $n \times p$ matrix $\mathbf{A}$ has the RIP if there exists a

restricted isometry constant δ_k and a *restricted orthogonality constant* $\theta_{k,k'}$ such that

$$\sqrt{1-\delta_K}\|\mathbf{c}\|_2 \leq \|\mathbf{A}\mathbf{c}\|_2 \leq \sqrt{1+\delta_k}\|\mathbf{c}\|_2 \quad \text{and} \quad |\mathbf{c}^T\mathbf{A}^T\mathbf{A}\mathbf{c}'| \leq \theta_{k,k'}\|\mathbf{c}\|_2\|\mathbf{c}'\|_2 \tag{1.9}$$

for all k-sparse $p \times 1$ vectors $\mathbf{c}$ and k'-sparse $p \times 1$ vectors $\mathbf{c}'$, where $\mathbf{c}$ and $\mathbf{c}'$ have disjoint supports. In the linear regression setting, a finite-sample error bound for the Dantzig selector was derived by assuming that the design matrix had the RIP, which amounts to assuming that the observed covariate vectors are not too collinear.

In the estimation equation setting, however, collinearity depends not only on the design matrix but also varies across the parameter space (Mackinnon and Puterman, 1989; Lesaffre and Marx, 1993), such that a model may be collinear even if its design matrix is not. Thus to derive a finite-sample bound for our NCOS, we assume that the conditions of the RIP hold across the entire parameter space $\boldsymbol{\Theta}$.

Assumption 3 *The estimating equation* $\mathbf{U}(\boldsymbol{\beta})$ *is differentiable with respect to* $\boldsymbol{\beta}$*. Let the negative Jacobian* $-\partial\mathbf{U}/\partial\boldsymbol{\beta}$ *be denoted* $\mathbf{A}(\boldsymbol{\beta})$*.*

Assumption 4 *There exist constants* $\delta_k > 0$ *and* $\theta_{k,k'}$ *such that for all* k*-sparse* $\mathbf{c}$ *and* k'*-sparse* $\mathbf{c}'$ *with disjoint supports,*

$$\delta_k\|\mathbf{c}\|_2^2 \leq |\mathbf{c}^T\mathbf{A}(\boldsymbol{\beta})\mathbf{c}| \quad \textit{and} \quad |\mathbf{c}^T\mathbf{A}(\boldsymbol{\beta})\mathbf{c}'| \leq \theta_{k,k'}\|\mathbf{c}\|_2\|\mathbf{c}'\|_2 \tag{1.10}$$

for all $\boldsymbol{\beta} \in \boldsymbol{\Theta}$*.*

When $\mathbf{U}(\boldsymbol{\beta})$ is the score function of a likelihood, $\mathbf{A}(\boldsymbol{\beta})$ from Assumption 3 is the observed information matrix evaluated at $\boldsymbol{\beta}$. Assumption 4 is similar to the RIP. The first part of Assumption 4 requires that each $k \times k$ submatrix of $\mathbf{A}(\boldsymbol{\beta})$ be invertible. If $\theta_{k,k'}$ is small, the second part of Assumption 4 roughly requires that each $k^* \times k^*$ submatrix act like a scalar multiple of an identity matrix, where $k^* = \max(k, k')$. If $\boldsymbol{\Theta}$ is unbounded, it is not clear

that δ_k and $\theta_{k,k'}$ will be finite. In the case of linear regression, $\mathbf{A}(\boldsymbol{\beta})$ is in fact independent of $\boldsymbol{\beta}$, so that the RIP constants are finite over $\boldsymbol{\Theta} = \mathbb{R}^p$. In general, however, since we do not expect the components of $\boldsymbol{\beta}$ to be arbitrarily large, we could assume that $\boldsymbol{\Theta}$ is bounded.

1.5.2 Decomposable norm-based regularizing functions

We will consider regularizers $r(\boldsymbol{\beta})$ that are norms and that are *decomposable*, a concept introduced by Negahban et al. (2009). Decomposability is defined relative to a subspace of A of $\mathbb{R}^p$, which is termed the *model subspace* and represents information about $\boldsymbol{\beta}_0$. For example, it may be the subspace of vectors defined by the non-zero coordinates of $\boldsymbol{\beta}_0$. Its orthogonal complement $A^\perp$ is termed the *perturbation subspace* and represents deviations away from $\boldsymbol{\beta}_0$. We decompose any $\boldsymbol{\beta}$ into its projections onto the model and perturbation subspaces, which we will denote $\mathbf{x}$ and $\mathbf{y}$, such that $\boldsymbol{\beta} = \mathbf{x} + \mathbf{y}$.

Negahban et al. (2009) defined $r(\boldsymbol{\beta})$ to be decomposable with respect to A if

$$r(\mathbf{x} + \mathbf{y}) = r(\mathbf{x}) + r(\mathbf{y}) \text{ for all } \mathbf{x} \in A, \mathbf{y} \in A^\perp. \tag{1.11}$$

To understand the intuitive rationale behind this property, consider that we would like to minimize $r(\mathbf{y})$, the norm of the components of $\boldsymbol{\beta}$ in the perturbation subspace. However, we generally do not know which components of $\boldsymbol{\beta}$ are in $A^\perp$, so we can only minimize the norm of the entire vector $\boldsymbol{\beta}$. But because r is a norm, $r(\boldsymbol{\beta}) \leq r(\mathbf{x}) + r(\mathbf{y})$, so minimizing $r(\boldsymbol{\beta})$ will not necessarily minimize $r(\mathbf{y})$ unless we have equality. This is exactly what is required by decomposability. Note that when we minimize a decomposable $r(\boldsymbol{\beta})$, we minimize $r(\mathbf{x})$ in addition to $r(\mathbf{y})$, so that our estimates for the components of $\boldsymbol{\beta}_0$ in the model subspace will be biased toward zero.

It turns out that many useful regularizing functions satisfy decomposability. If $r(\boldsymbol{\beta})$ is the ℓ_1-norm, then consider $T_0 = \{j : \beta_{0j} \neq 0\}$ and $A_0 = \{\mathbf{x} \in \mathbb{R}^p : x_j \neq 0 \text{ for } j \in T_0\}$. It is clear that the ℓ_1-norm is decomposable with respect to A_0. Next, suppose $\boldsymbol{\beta}_0$ obeys

some group structure, and let T_0 be such that T_0^c contains the indices corresponding to the unimportant groups. If $A_0 = \{\mathbf{x} \in \mathbb{R}^p : x_j = 0 \text{ for } j \notin T_0\}$, then the group lasso penalty, which is a sum of ℓ_2-norms and is therefore a norm, is decomposable with respect to A_0. In addition, it is easy to see that the adaptive versions of these regularizers, where each component of $\boldsymbol{\beta}$ is divided by a consistent initial estimator, are also decomposable norms. More details are given in Negahban et al. (2009).

1.5.3 Error bound

With these assumptions on $\mathbf{U}(\boldsymbol{\beta})$ and $r(\boldsymbol{\beta})$, we can state a finite-sample bound on the error of our NCOS estimator. We first define a few terms. With A_0 and T_0 defined as in Section 1.5.2, consider a regularizing function $r(\boldsymbol{\beta})$ that is a decomposable norm with respect to A_0, and let $|T_o| = k$. To bound the ℓ_2-norm of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$, we will need to link the regularizing norm to the ℓ_2-norm. To this end we use the concept of *subspace compatibility constants*, introduced by Negahban et al. (2009), which is related to the topological concept of equivalent norms. Let $\Psi_{12}(A_0)$, Ψ, and $\Psi_{r2}(A_0)$ be three subspace compatibility constants, which are defined in greater detail in Appendices 1.7.1 and 1.7.2.

Theorem 1 *Let $\mathbf{U}(\boldsymbol{\beta})$ be an estimating function satisfying Assumptions 1–4, and assume that $\boldsymbol{\beta}_0$ is k-sparse. If*

$$c_1 = \delta_{1.25k} - \theta_{k,1.25k}\Psi_{12}(A_0)^{-1}\Psi\Psi_{r2}(A_0) > 0, \tag{1.12}$$

then the NCOS estimate $\hat{\boldsymbol{\beta}}$ obtained by solving (1.1) obeys

$$\mathrm{P}\left\{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq \frac{(10k)^{1/2}\gamma}{c_1}\right\} \geq 1 - 2p\exp\left\{-\frac{n\gamma^2}{2(v + M\gamma)}\right\}, \tag{1.13}$$

where $\gamma > 0$ is nonrandom tuning parameter and v and M are positive constants defined in Assumption 2.

For any n and p, Theorem 1 gives an upper bound for the mean squared error of $\hat{\boldsymbol{\beta}}$ that holds with high probability. The requirement that $\boldsymbol{\beta}_0$ be sparse is common in high-dimensional data analysis, and without it, it is very difficult to get useful bounds on estimation error. The size of our bound depends on the subspace compatibility constants, and when $r(\boldsymbol{\beta})$ is the ℓ_1-norm or the group lasso penalty, Negahban et al. (2009) showed that the Ψ terms are finite. Theorem 1 is similar to the error bounds of the Dantzig selector in the linear model (Candès and Tao, 2007) and the Cox model (Antoniadis et al., 2010). In those cases, the estimator is capable of achieving, with high probability, a rate of convergence within a factor of $\log p$ of the optimal rate that an oracle estimator would provide.

The condition on c_1 roughly requires that every submatrix of $\mathbf{A}(\boldsymbol{\beta})$ with at most $1.25k$ columns be approximately invertible and orthogonal for all $\boldsymbol{\beta} \in \boldsymbol{\Theta}$. Similar assumptions were made by Candès and Tao (2007), Antoniadis et al. (2010), and Cai et al. (2010). In particular, if $r(\boldsymbol{\beta})$ is the ℓ_1-norm, then $\Psi_{12}(A_0)^{-1}\Psi\Psi_{r2}(A_0) = 1$ (see Appendix) and our condition on c_1 reduces to $\delta_{1.25k} > \theta_{k,1.25k}$, which is weaker than the corresponding conditions of Candès and Tao (2007) and Antoniadis et al. (2010). If $r(\boldsymbol{\beta})$ is the group lasso regularizer, then $\Psi_{12}(A_0)^{-1}\Psi\Psi_{r2}(A_0)$ equals the square root of the number of groups with nonzero components, multiplied by the maximum group size, divided by k. This is again close to 1 if groups have roughly the same size.

1.6 Discussion

In this paper we have proposed a new sparse estimation procedure for estimating equations. We have shown that the NCOS can give good results in real data and in simulations, performing simultaneous estimation and variable selection with lasso and group lasso penalties for two different estimating equations. Our implementation of the NCOS, using a sequential linear programming strategy with a trust region and a filter, also enjoys global convergence properties (Huang et al., 2011). Finally, for decomposable norm-based regularizers, which

include the lasso and the group lasso, we have provided a probability bound on the ℓ_2-error of the NCOS parameter estimates.

While our Theorem 1 applies only to a certain class of estimating equations and a certain class of regularizers, the NCOS algorithm itself does not require these restrictions. In fact, our sequential linear programming implementation can be used with a wide variety of estimating equations and $r(\boldsymbol{\beta})$, including nonconvex $r(\boldsymbol{\beta})$ such as SCAD (Fan and Li, 2001) or the group bridge penalty for between- and within-group selection (Huang et al., 2009; Wang et al., 2009). While it is difficult to derive finite-sample error bounds for these $r(\boldsymbol{\beta})$, it would be interesting to investigate their asymptotic properties, along the lines of Dicker (2011).

In this work we have focused on smooth estimating equations, but in some cases $\mathbf{U}(\boldsymbol{\beta})$ may not be differentiable. In this situation the implementation of the NCOS would require a different algorithm, such as derivative-free constrained optimization (Conn et al., 2009). So far, however, these methods are limited in the number of variables they can reasonably accommodate. Nevertheless, we believe that our NCOS is a flexible and useful strategy for regularizing estimating equations.

1.7 Appendix A: Proof of Theorem 1

1.7.1 Subspace compatibility constant

We will need to define the *subspace compatibility constant*, introduced by Negahban et al. (2009). For any pair of norms a and b and subspace A, the subspace compatibility constant is

$$\Psi_{ab}(A) = \inf\{c > 0 | a(u) \leq cb(u) \text{ for all } u \in A\}, \tag{1.14}$$

and is a measure of how similar the norms a and b are over the subset A. Below, we will denote the ℓ_1-, ℓ_2-, and r- norms by replacing a or b in Ψ_{ab} with 1, 2, or r. For example, if

A has dimension k, then $\Psi_{12}(A) = k^{1/2}$.

1.7.2 Proof

Recall the set of indices T_0 from Section 1.5.2. Let $\mathbf{h}$ be the vector constructed by arranging the components of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ in decreasing order of magnitude after placing the elements of T_0 first, and assume that the entries of $\mathbf{U}(\boldsymbol{\beta})$ and $\mathbf{A}(\boldsymbol{\beta})$ are reordered accordingly. Relabel $T_0 = \{1, \ldots, k\}$, and let $T_* = \{k+1, \ldots, 5k/4\}$, and $T_i, i \geq 1$ be successive subsets, each of size k, of the remaining indices $\{5k/4 + 1, \ldots, p\}$. If $k < 4$, then $T_* = \{k+1\}$, and the last T_i can contain fewer than k elements. For sets T_1 and T_2, let $\mathbf{c}_{T_1}$ denote the subvector of $\mathbf{c}$ consisting of the components indexed by T_1, and $\mathbf{X}_{T_1,T_2}$ denotes the submatrix of $\mathbf{X}$ consisting of the rows indexed by T_1 and the columns indexed by T_2. Let $\mathbf{X}_{T_1,\cdot}$ denote the submatrix with rows indexed by T_1 and containing all p columns. Finally, redefine the subspace A_0 as in Section 1.5.2 and define A_* and A_i analogously.

We first note that by the fundamental theorem of calculus, for the j^{th} component U_j of $\mathbf{U}$,

$$U_j(\hat{\boldsymbol{\beta}}) - U_j(\boldsymbol{\beta}_0) = -\int_0^1 \sum_{l=1}^{p} A_{jl}(\boldsymbol{\beta}_0 + t\mathbf{h}) h_l dt, \tag{1.15}$$

where A_{jl} is the jl^{th} element of the negative Jacobian $\mathbf{A}$ and h_l is the l^{th} component of $\mathbf{h}$. Define $\mathbf{D}$ to be the $p \times p$ matrix where the jl^{th} element

$$D_{jl} = \int_0^1 A_{jl}(\boldsymbol{\beta}_0 + t\mathbf{h}) dt. \tag{1.16}$$

First, by the Cauchy-Schwarz inequality,

$$\left|\mathbf{h}^T_{T_0 \cup T_*,\cdot} \mathbf{D}_{T_0 \cup T_*,\cdot} \mathbf{h}\right| \leq \|\mathbf{h}_{T_0 \cup T_*}\|_2 \|\mathbf{D}_{T_0 \cup T_*,\cdot} \mathbf{h}\| \tag{1.17}$$

$$= \|\mathbf{h}_{T_0 \cup T_*}\|_2 \|\mathbf{U}(\hat{\boldsymbol{\beta}})_{T_0 \cup T_*} - \mathbf{U}(\boldsymbol{\beta}_0)_{T_0 \cup T_*}\|_2. \tag{1.18}$$

By the definition of the NCOS optimization problem (1.1), $\|\mathbf{U}(\hat{\boldsymbol{\beta}})\|_\infty \leq \gamma$. We will also later

show that $\|\mathbf{U}(\boldsymbol{\beta}_0)\|_\infty \leq \gamma$ with high probability. We can therefore conclude that

$$\left|\mathbf{h}^T_{T_0\cup T_*}\mathbf{D}_{T_0\cup T_*,\cdot}\mathbf{h}\right| \leq \|\mathbf{h}_{T_0\cup T_*}\|_2(1.25k)^{1/2}2\gamma. \tag{1.19}$$

Now, note that for any k-sparse vector $\mathbf{c}$,

$$|\mathbf{c}^T\mathbf{D}\mathbf{c}| = \left|\int_0^1 \sum_{jl} c_j A_{jl}(\boldsymbol{\beta}_0 + t\mathbf{h})c_l dt\right| \geq \left|\int_0^1 \delta_k\|\mathbf{c}\|_2^2 dt\right| = \delta_k\|\mathbf{c}\|_2^2, \tag{1.20}$$

because by Assumption 4 the restricted isometry constant δ_k of $\mathbf{A}(\boldsymbol{\beta})$ holds for all $\boldsymbol{\beta} \in \boldsymbol{\Theta}$. Similarly, for any k'-sparse $\mathbf{c}'$ where $\mathbf{c}$ and $\mathbf{c}'$ have disjoint support, we find that

$$|\mathbf{c}^T\mathbf{D}\mathbf{c}'| \leq \theta_{k,k'}\|\mathbf{c}\|_2\|\mathbf{c}'\|_2. \tag{1.21}$$

Therefore,

$$\begin{aligned}
\left|\mathbf{h}^T_{T_0\cup T_*}\mathbf{D}_{T_0\cup T_*,\cdot}\mathbf{h}\right| &= \left|\mathbf{h}^T_{T_0\cup T_*}\mathbf{D}_{T_0\cup T_*,T_0\cup T_*}\mathbf{h}_{T_0\cup T_*} + \sum_{i\geq 1}\mathbf{h}^T_{T_0\cup T_*}\mathbf{D}_{T_0\cup T_*,T_i}\mathbf{h}_{T_i}\right| && (1.22)\\
&\geq \left|\mathbf{h}^T_{T_0\cup T_*}\mathbf{D}_{T_0\cup T_*,T_0\cup T_*}\mathbf{h}_{T_0\cup T_*}\right| - \sum_{i\geq 1}\left|\mathbf{h}^T_{T_0\cup T_*}\mathbf{D}_{T_0\cup T_*,T_i}\mathbf{h}_{T_i}\right| && (1.23)\\
&\geq \delta_{1.25k}\|\mathbf{h}_{T_0\cup T_*}\|_2^2 - \sum_{i\geq 1}\theta_{k,1.25k}\|\mathbf{h}_{T_0\cup T_*}\|_2\|\mathbf{h}_{T_i}\|_2. && (1.24)
\end{aligned}$$

We focus on the $\sum_{i\geq 1}\|\mathbf{h}_{T_i}\|_2$. If we denote the first $k/4$ elements of T_i by T_{i1} and the last $3k/4$ elements by T_{i2}, and define the subspaces A_{i1} and A_{i2} accordingly, then using the shifting inequality of Cai et al. (2010) we can show that

$$\begin{aligned}
\|\mathbf{h}_{T_1}\|_2 &\leq k^{-1/2}\left(\|\mathbf{h}_{T_*}\|_1 + \|\mathbf{h}_{T_{11}}\|_1\right) && (1.25)\\
&\leq \Psi_{12}(A_0)^{-1}\left\{\Psi_{1r}(A_*)r(\mathbf{h}_{T_*}) + \Psi_{1r}(A_{11})r(\mathbf{h}_{T_{11}})\right\}, && (1.26)\\
\|\mathbf{h}_{T_i}\|_2 &\leq k^{-1/2}\left(\|\mathbf{h}_{T_{(i-1)2}}\|_1 + \|\mathbf{h}_{T_{i1}}\|_1\right) && (1.27)\\
&\leq \Psi_{12}(A_0)^{-1}\left\{\Psi_{1r}(A_{(i-1)2})r(\mathbf{h}_{T_{(i-1)2}}) + \Psi_{1r}(A_{i1})r(\mathbf{h}_{T_{i1}})\right\}, i = 2,3,\ldots && (1.28)
\end{aligned}$$

Then we can conclude that $\sum_{i\geq 1}\|\mathbf{h}_{T_i}\|_2 \leq \Psi_{12}(A_0)^{-1}\Psi r(\mathbf{h}_{T_0^c})$, where

$$\Psi = \max\{\Psi_{1r}(A_*), \Psi_{1r}(A_{i1}), \Psi_{1r}(A_{i2}), i = 1,2,\ldots\}. \tag{1.29}$$

Next, by the definition of our optimization problem, if $\boldsymbol{\beta}_0$ is feasible then we must have that $r(\boldsymbol{\beta}_0) \geq r(\hat{\boldsymbol{\beta}})$ (we will show later that $\boldsymbol{\beta}_0$ is indeed feasible with high probability). Also, by assumption $r(\boldsymbol{\beta})$ is decomposable with respect to A_0, so $r(\hat{\boldsymbol{\beta}}) = r(\hat{\boldsymbol{\beta}}_{T_0}) + r(\hat{\boldsymbol{\beta}}_{T_0^c})$. Combining these facts and using the triangle inequality, we find that $r(\boldsymbol{\beta}_0) \geq r(\boldsymbol{\beta}_{0T_0}) - r(\mathbf{h}_{T_0}) + r(\hat{\boldsymbol{\beta}}_{T_0^c})$. Since $\boldsymbol{\beta}_{0T_0} = \boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_{0T_0^c} = \mathbf{0}$, we see that $r(\mathbf{h}_{T_0}) \geq r(\hat{\boldsymbol{\beta}}_{T_0^c}) = r(\hat{\boldsymbol{\beta}}_{T_0^c} - \boldsymbol{\beta}_{0T_0^c}) = r(\mathbf{h}_{T_0^c})$.

We can therefore conclude that

$$\sum_{i \geq 1} \|\mathbf{h}_{T_i}\|_2 \leq \Psi_{12}(A_0)^{-1} \Psi r(\mathbf{h}_{T_0}) \leq \Psi_{12}(A_0)^{-1} \Psi \Psi_{r2}(A_0) \|\mathbf{h}_{T_0}\|_2 \tag{1.30}$$

$$\leq \Psi_{12}(A_0)^{-1} \Psi \Psi_{r2}(A_0) \|\mathbf{h}_{T_0 \cup T_*}\|_2. \tag{1.31}$$

By the triangle inequality, $\|\mathbf{h}_{T_0^c}\|_2 \leq \sum_{i \geq 1} \|\mathbf{h}_{T_i}\|_2$, so

$$\left|\mathbf{h}_{T_0 \cup T_*}^T \mathbf{D}_{T_0 \cup T_*, \cdot} \mathbf{h}\right| \geq \left\{\delta_{1.25k} - \theta_{k,1.25k} \Psi_{12}(A_0)^{-1} \Psi \Psi_{r2}(A_0)\right\} \|\mathbf{h}_{T_0 \cup T_*}\|_2^2. \tag{1.32}$$

Because $\|\mathbf{h}\|_2^2 = \|\mathbf{h}_{T_0 \cup T_i}\|_2^2 + \sum_{i \geq 1} \|\mathbf{h}_{T_i}\|_2^2 \leq \|\mathbf{h}_{T_0 \cup T_i}\|_2^2 + (\sum_{i \geq 1} \|\mathbf{h}_{T_i}\|_2)^2 \leq 2\|\mathbf{h}_{T_0 \cup T_i}\|_2^2$, we can combine (1.19) and (1.32) to find that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq (10k)^{1/2}\gamma / c_1$.

Finally, we show that $\|\mathbf{U}(\boldsymbol{\beta}_0)\| \leq \gamma$ with high probability, which implies that $\boldsymbol{\beta}_0$ is feasible. But by Assumptions 1 and 2 and Bernstein's inequality,

$$\mathrm{P}(\|\mathbf{U}(\boldsymbol{\beta}_0)\|_\infty \geq \gamma) \leq 2p \exp\left\{-\frac{n^2\gamma^2}{2(nv + Mn\gamma)}\right\}, \tag{1.33}$$

and we have proven Theorem 1.

Principled sure independence screening for Cox models with ultra-high-dimensional covariates

Sihai Dave Zhao and Yi Li

Department of Biostatistics
Harvard School of Public Health

2.1 Introduction

An urgent need has emerged in the field of biomedicine for statistical procedures capable of analyzing and interpreting vast quantities of data. Selecting the best predictors of an outcome is a key step in this process, but traditional methods of variable selection, such as best subset selection or backward selection, have been found to be unstable and inaccurate when the dimension of the covariates is close to the number of observations. Furthermore, when there are more covariates than observations, as is often the case in genomic studies, these methods can fail completely.

To address these issues, recent work has focused on regularized regression procedures such as the lasso (Tibshirani, 1996) and adaptive lasso (Zou, 2006), the elastic net (Zou and Hastie, 2005), the smoothly clipped absolute deviation estimator (Fan and Li, 2001), and the Dantzig selector (Candès and Tao, 2007). These methods can handle the high-dimension-low-sample-size paradigm, have superior predictive accuracy, and under certain conditions can achieve the oracle property (Fan and Li, 2001): they are as accurate and efficient as an estimator that knows *a priori* which variables are truly important.

However, these procedures only work well with a moderate number of covariates. When the dimension of the covariates is ultra-high, both traditional and regularization methods have problems with speed, stability, and accuracy (Fan and Lv, 2008). For example, many of the bounds on the accuracy of these methods involve factors of $\log p_n$, where p_n is the dimension of the covariates (Candès and Tao, 2007; Wainwright, 2009). Thus the theoretical performance of these methods degrades as p_n becomes very large, yet this ultra-high dimensionality characterizes many real-world biological datasets. Our work in this paper is motivated by one such dataset, in the area of multiple myeloma.

Multiple myeloma is the world's second-most common hematological cancer and patients often present with bone lesions, immunological disorders, and renal failure. An effective

treatment is still being sought, as only about 10% of patients survive 10 years after diagnosis. A deeper understanding of the molecular etiology of this disease would lead to novel therapeutic targets and more accurate risk classification systems. We studied overall survival for 80 multiple myeloma patients enrolled in a clinical trial of bortezomib (Mulligan et al., 2007). With expression level measurements on 44760 probesets, this dataset defies analysis even with regularized regression.

Without tools to deal with this type of ultra-high dimensionality, many analysts employ an initial univariate screening step to reduce the number of covariates under consideration. The remaining covariates could then be fed to one of the more sophisticated regularization techniques in a second stage. But it was only recently that Fan and Lv (2008) placed this ad-hoc practice on firm theoretical ground, showing that that screening could indeed improve the performance of regularization methods. They suggested fitting marginal regression models for each covariate, choosing a threshold, and retaining those covariates for which the magnitudes of the parameter estimates are above the threshold. When the data come from an ordinary linear model with normal errors, Fan and Lv (2008) showed that this pre-screening procedure, which they termed sure independence screening (SIS), has desirable theoretical properties. Fan and Song (2010) later gave theoretical justification for using SIS with generalized linear models.

But two important problems remain. First, one common type of outcome data seen in clinical settings, including in our myeloma dataset, is survival time, which is subject to censoring. Regularized regression methods for censored observations have been studied, as reviewed in Li (2008), but these are subject to the same issues mentioned above when the dimension of the covariates is ultra-high. There is thus a need for a pre-screening procedure in this setting, but the results of Fan and Lv (2008) and Fan and Song (2010) cannot be applied because the issue of censoring is not addressed. Several ad-hoc solutions are available from Tibshirani (2009) and Fan et al. (2010), but none of these proposals has much theoretical support. The extension of the theoretical sure screening results to censored data is not immediate because it turns out that certain conditions on the relationship between the

covariates and the censoring distribution are required for screening to have good theoretical properties, an issue which does not emerge with uncensored data.

The second problem is that existing screening procedures require choosing a threshold to dictate how many variables to retain, but there are no principled methods for making such a choice, making the resulting screened models difficult to evaluate. The threshold can be thought of as a regularization parameter, which in the regression setting is ordinarily chosen by optimizing out-of-sample prediction error using cross-validation or generalized cross-validation. However, this approach is unavailable for screening procedures because no prediction rule is ever generated.

In this paper we provide a screening method for censored survival data with ultra-high-dimensional covariates. We also propose a new, principled method for choosing the number of covariates to retain based on specifying the desired false positive rate. Finally, we give, to our knowledge, the first theoretical justifications of the sure independence screening procedure for censored data. Under the asymptotic framework where the number of covariates can grow with the sample size, we show that with probability going to 1, our procedure will select all of the important variables with a false positive rate close to the prespecified level.

Our paper is organized as follows. We briefly review sure independence screening for generalized linear models in Section 2.2. In Section 2.3 we discuss the implementation and the theoretical properties of our principled sure independence screening procedure, and present simulation results in Section 2.4. Section 2.5 describes our analysis of the myeloma dataset, and we conclude with a discussion in Section 2.6. All proofs are given in the Appendix.

2.2 Sure independence screening in generalized linear models

We first review the sure independence screening formulation of Fan and Song (2010). For subjects $i = 1, \ldots, n$ let $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip_n})$ be the p_n-dimensional covariate vector. Assuming that observations Y_i come from an exponential family, we model $E(Y_i \mid \mathbf{Z}_i)$ as some function of a linear predictor $\boldsymbol{\alpha}_0^T \mathbf{Z}_i$ with parameter vector $\boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0p_n})$. When p_n is much larger than n we are unable to estimate $\boldsymbol{\alpha}_0$ with conventional procedures. To reduce p_n, sure independence screening proceeds by regressing Y_i on each Z_{ij} individually to calculate marginal maximum likelihood estimates $\hat{\beta}_j$. The final screened model retains all covariates $j : |\hat{\beta}_j| \geq \gamma_n$ for some prespecified constant cutoff γ_n.

Fan and Song (2010) showed that under certain conditions, if γ_n follows an ideal rate, this procedure has two desirable properties, namely the sure screening property and the size control property. The former guarantees that the screened model will contain the true model with a probability approaching 1. The latter states that if $\log(p_n) = o(n^{1-2\kappa})$ where $\kappa < 1/2$, the probability that the size of the screened model will be at most $O\{n^{2\kappa}\lambda_{max}(\boldsymbol{\Sigma})\}$ will also go to 1, where $\boldsymbol{\Sigma} = \text{var}(\mathbf{Z}_i)$ and $\lambda_{max}(\boldsymbol{\Sigma})$ is the largest eigenvalue of $\boldsymbol{\Sigma}$.

These results, however, are restricted to non-censored generalized linear models. Furthermore, it is difficult to translate the ideal rate for γ_n into a method for selecting the cutoff in practice. Fan and Lv (2008) suggest $n/log(n)$ or $n - 1$ as the number of covariates to retain after screening, but without theoretical justification. To address these issues, we investigate here a reliable pre-screening procedure in a survival setting, where the outcomes are subject to right censoring, and propose a principled method for choosing γ_n based on controlling the false positive rate.

2.3 Principled Cox sure independence screening

2.3.1 Method

In the context of survival analysis, we assume that the underlying survival times T_i follow a Cox model (Cox, 1972) with the true hazard function

$$\lambda(x; \mathbf{Z}_i) = \lambda_0(x) \exp(\boldsymbol{\alpha}_0^T \mathbf{Z}_i), \tag{2.1}$$

where $\lambda_0(x)$ is unspecified. Let $\tilde{C}_i$ be potential censoring times, which are independent of T_i conditional on $\mathbf{Z}_i$. Furthermore let $\tau > 0$ be the finite study duration such that $\mathrm{P}\{\min(\tilde{C}_i, \tau) < T_i\} < 1$, ensuring that enough events will be observed over $[0, \tau]$. The effective censoring times are thus $C_i = \min(\tilde{C}_i, \tau)$. We observe $X_i = \min(T_i, C_i)$, and $\delta_i = I(T_i \leq C_i)$. Without loss of generality, we assume throughout that $E(Z_{ij}) = 0$ for all j.

To perform an initial screening procedure, we propose to fit marginal Cox regressions, possibly misspecified, for each Z_{ij}, namely $\lambda_0^*(x) \exp(\beta Z_{ij})$. Let $N_i(t) = I(X_i \leq t, \delta_i = 1)$ be independent counting processes for each subject i and $Y_i(t) = I(X_i \geq t)$ be the at-risk processes. For $k = 0, 1, \ldots$, define

$$S_j^{(k)}(x) = n^{-1} \sum_{i=1}^{n} Z_{ij}^k Y_i(x) \lambda(x; \mathbf{Z}_i), \quad s_j^{(k)}(x) = E\{S^{(k)}(x)\}, \tag{2.2}$$

$$S_j^{(k)}(\beta, x) = n^{-1} \sum_{i=1}^{n} Z_{ij}^k Y_i(x) \exp(\beta Z_{ij}), \quad s_j^{(k)}(\beta, x) = E\{S^{(k)}(\beta, x)\}. \tag{2.3}$$

Then the maximum marginal partial likelihood estimator $\hat{\beta}_j$ solves the estimating equation

$$U_j(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_{ij} - \frac{S_j^{(1)}(\beta, x)}{S_j^{(0)}(\beta, x)} \right\} dN_i(x) = 0. \tag{2.4}$$

Finally, let β_{0j} be the solution to the limiting estimation equation

$$u_j(\beta) = \int_0^{\tau} \left\{ s_j^{(1)}(x) - \frac{s_j^{(1)}(\beta, x)}{s_j^{(0)}(\beta, x)} s_j^{(0)}(x) \right\} dx. \tag{2.5}$$

Define the information matrix to be $I_j(\beta) = -\partial U_j/\partial\beta$ at $\hat{\beta}_j$. We will denote the final screened model by $\hat{\mathcal{M}} = \{j : I_j(\hat{\beta}_j)^{1/2}|\hat{\beta}_j| \geq \gamma_n\}$. We would like a practical way of choosing γ_n such that we can achieve the sure screening property while controlling the false positive rate, or the proportion of unimportant covariates we incorrectly include in $\hat{\mathcal{M}}$. If the true model $\mathcal{M} = \{j : \alpha_{0j} \neq 0\}$ has size $|\mathcal{M}| = s_n$, then the expected false positive rate can be written as

$$\mathrm{E}\left(\frac{\left|\hat{\mathcal{M}} \cap \mathcal{M}^c\right|}{|\mathcal{M}^c|}\right) = \frac{1}{p_n - s_n} \sum_{j \in \mathcal{M}^c} \mathrm{P}\left\{I_j(\hat{\beta}_j)^{1/2}|\hat{\beta}_j| \geq \gamma_n\right\}. \tag{2.6}$$

We can show that $I_j(\hat{\beta}_j)^{1/2}\hat{\beta}_j$ has an asymptotically standard normal distribution, so we see that γ_n corresponds to controlling the expected false positive rate at $2\{1 - \Phi(\gamma_n)\}$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

However, we would like the false positive rate to decrease to 0 as p_n increases with n, though it can never exactly equal 0 or else $\gamma_n = \infty$. One sensible way to do this would be to first fix the number of false positives f that we are willing to tolerate, which would correspond to a false positive rate of $f/(p_n - s_n)$. Because s_n is unknown, we can be conservative by letting $\gamma_n = \Phi^{-1}\{1 - q_n/2)\}$ where $q_n = f/p_n$, so that the expected false positive rate is $2\{1 - \Phi(\gamma_n)\} = q_n \leq f/(p_n - s_n)$. We can show that this procedure maintains the sure screening property, and more precise arguments will be given later (Theorems 5 and 6).

We term this method a principled Cox sure independence screening procedure (abbreviated PSIS), as the cutoff γ_n is selected to control the false positive rate. Specifically, PSIS is implemented as follows:

1. Fit a marginal Cox model for each of the covariates according to equation (2.4) to get parameter estimates $\hat{\beta}_j$ and variance estimates $I_j(\hat{\beta}_j)^{-1}$.
2. Fix the false positive rate $q_n = f/p_n$ and let $\gamma_n = \Phi^{-1}(1 - q_n/2)$.
3. Retain covariates $j : I_j(\hat{\beta}_j)^{1/2}|\hat{\beta}_j| \geq \gamma_n$.

Our cutoff selection procedure is related to false discovery rate (FDR) methods (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). In particular, the FDR is defined as $|\hat{\mathcal{M}} \cap \mathcal{M}^c|/|\hat{\mathcal{M}}|$, which is simply the product of the false positive rate in (2.6) and $|\mathcal{M}^c|/|\hat{\mathcal{M}}|$, which is less than $p_n/|\hat{\mathcal{M}}|$. Therefore, controlling the false positive rate at $q_n = f/p_n$ is equivalent to controlling the FDR at $f/|\hat{\mathcal{M}}|$, conditional on $|\hat{\mathcal{M}}|$. Bunea et al. (2006) have in fact shown that FDR methods can also have the sure screening property, though only in the linear regression case.

Our screening procedure resembles the "marginal ranking" methods for censored outcome data proposed by various authors (Fan et al., 2010; Tibshirani, 2009). However, to our knowledge, none of these proposals has much theoretical support. A much more aggressive method of control has been proposed by Fan et al. (2010). We show below that our proposed procedure maintains the sure screening property, and will also control the false positive rate at close to the nominal level. Fan and Lv (2008) also proposed an iterative sure independence screening procedure (ISIS) for linear models, which they showed can perform better than SIS. However, they were unable to offer theoretical support. In this paper we focus on first understanding non-iterative screening for the Cox model.

2.3.2 Theoretical properties

First, under certain assumptions, we find that we can distinguish α_{0j}, $j \in \mathcal{M}$ from α_{0j}, $j \in \mathcal{M}^c$ in the presence of censoring. It is this guarantee that makes the marginal screening approach possible.

Theorem 2 *Under Assumptions 5–12 in the Appendix, $\beta_{0j} = 0$ if and only if $\alpha_{0j} = 0$, for all $j = 1, \ldots, p_n$.*

Following Struthers and Kalbfleisch (1986) and under Assumptions 5 and 6 in the Appendix, we know that the $\hat{\beta}_j$ are consistent for β_{0j}. It is therefore natural to ask how accurate these estimates are.

Theorem 3 *Under Assumptions 5–12 in the Appendix,*

$$\mathrm{P}\left\{\sqrt{n}|\hat{\beta}_j - \beta_{0j}| \geq 4K[1 + \Lambda_0(\tau)\exp\{2K(A+L)\}](1+t)/H\right\} \leq \exp(-t^2/2) \tag{2.7}$$

for all $j = 1, \ldots, p_n$, where K is the bound on the covariates Z_{ij} for all j, $\Lambda_0(\tau) = \int_0^\tau \lambda_0(s)ds$ is bounded by Assumption 8, A is the bound on the parameters α_{0j} for all j, $L = \|\boldsymbol{\alpha}_0\|_1$ and is bounded by Assumption 7, and H is defined in Assumption 9.

Theorem 3 is important as it suggests that $|\hat{\beta}_j - \beta_{0j}|$ is at most on the order of $n^{-1/2}$ with high probability. Hence in order to detect covariate $j \in \mathcal{M}$, we need $|\beta_{0j}|$ to be at least $O(n^{-1/2})$, which is indeed the case as shown by the following theorem.

Theorem 4 *Under Assumptions 5–12 in the Appendix, there is a constant $c_2 > 0$ such that $\min_{j\in\mathcal{M}} |\beta_{0j}| \geq c_2 n^{-\kappa}$, where $\kappa < 1/2$.*

Because the $|\beta_{0j}|$ are large enough to be detected with our marginal Cox regressions, and because they reflect the importance of the Z_{ij} in the true joint model, we can prove that our procedure maintains the sure screening property and controls the false positive rate at close to the nominal level.

Theorem 5 (Sure screening property) *Under Assumptions 5–12 in the Appendix, if we choose $\gamma_n = \Phi^{-1}(1-q_n/2)$, then for $\kappa < 1/2$ and $\log(p_n) = O(n^{1/2-\kappa})$, there exists a constant $c_3 > 0$ such that*

$$\mathrm{P}(\mathcal{M} \subseteq \hat{\mathcal{M}}) \geq 1 - s_n \exp(-c_3 n^{1-2\kappa}). \tag{2.8}$$

Theorem 6 (False positive control property) *Under Assumptions 5–12 in the Appendix, if we choose* $\gamma_n = \Phi^{-1}(1 - q_n/2)$*, then there exists some* $c_4 > 0$ *such that*

$$\mathrm{E}\left(\frac{|\hat{\mathcal{M}} \cap \mathcal{M}^c|}{|\mathcal{M}^c|}\right) \leq q_n + c_4 n^{-1/2}, \tag{2.9}$$

where $|\hat{\mathcal{M}} \cap \mathcal{M}^c|/|\mathcal{M}^c|$ *can be interpreted as the false positive rate.*

It is often assumed that the true model is sparse and s_n is small (Candès and Tao, 2007), in which case Theorem 5 indicates that we will be able to retain all important covariates with high probability. The probability bound will converge to 1 if $\log(p_n) = O(n^{1/2-\kappa})$, which is comparable to the rates allowed in Fan and Lv (2008) and Fan and Song (2010). That p_n is allowed to incease exponentially justifies the use of sure independence screening in the Cox model when p_n is ultra-high-dimensional.

2.4 Simulations

To evaluate the finite-sample performance of our sure screening and false positive control properties, we performed PSIS on simulated datasets generated from Cox models and examined its average false positive and negative rates. We simulated 200 datasets, each consisting of $p_n = 20000$ covariates and $n = 100$ subjects. We generated the covariates from a multivariate normal distribution where the mean was 0 and the correlation between components Z_{ij} and Z_{ik} was $\rho^{|j-k|}$ for $\rho = 0.5$, and 0.9. We next generated survival times from Cox models with baseline hazards of $\lambda_0(x) = 1$ and linear predictors $\boldsymbol{\alpha}_0^T \mathbf{Z}_i$ for different parameter vectors $\boldsymbol{\alpha}_0$. We let the number of non-zero elements of $\boldsymbol{\alpha}_0$ be either $s_n = 10$ or 20 and set the first s_n components of $\boldsymbol{\alpha}_0$ to be either all equal to 0.35 or all equal to 0.7. Finally, we generated censoring times from a uniform and an exponential distribution, which gave bounded and unbounded censoring times respectively. Under each censoring mechanism, we considered censoring rates of approximately 20%, 50%, and 70%.

To explore how the variable selection performance of a few popular regularized regression techniques were affected by PSIS with different values of q_n, we followed PSIS by either lasso (Tibshirani, 1997), adaptive lasso (Zhang and Lu, 2007), or SCAD (Fan and Li, 2002). Since the initial parameter estimates required by adaptive lasso do not exist when $p_n > n$, we first applied ordinary lasso to reduce p_n and calculated the initial estimates using the remaining covariates. We implemented lasso and adaptive lasso using a coordinate descent algorithm (Friedman et al., 2007) with the R package `glmnet`, and we implemented SCAD using the one-step estimator of Zou and Li (2008) with the package `SIS`. We tuned each regularized regression with BIC to achieve selection consistency, and we denote these two-stage procedures by PSIS-L, PSIS-L-A, and PSIS-S, respectively.

Tables 2.1 and 2.2 report numerical results for our sure screening and false positive control properties when $s_n = 20$, $\alpha_{0j} = 0.35$, and when censoring times were generated from an exponential distribution. The results when $s_n = 10$, $\alpha_{0j} = 0.7$, or when the censoring times were uniformly distributed are similar and are omitted for the sake of space. We considered $q_n = 10^r$ for $r = -6, \ldots, -2$, and also ranging from 0.1 to 1 (corresponding to no screening) in increments of 0.1. The results support our principled cutoff procedure: the observed false positive rates closely match the nominal q_n when $\rho = 0.5$ for all censoring rates. When $\rho = 0.9$, the observed false positive rates can be higher than the nominal q_n for $q_n \leq 10^{-4}$, but since $p_n = 20000$ here, $q_n = 10^{-4}$ corresponds to only 2 false positives. In other words, even when the nominal q_n underestimates the true false positive rate, the absolute number of false positives selected by PSIS is still fairly small.

Figure 2.1 plots the average false negative rates for PSIS against q_n, which increase as q_n decreases but generally don't rise dramatically until $q_n \approx 0.1$. For a given q_n the false negative rates decrease with larger α_{0j}. The performance of PSIS actually improves when $\rho = 0.9$, perhaps because in our simulated data the correlation between the important covariates increases with ρ, making the marginal $\hat{\beta}_j$ estimates for those covariates more likely to be similar in magnitude. Higher censoring rates exhibit worse performance, as expected. These results suggest that q_n can be set fairly low and the false negative rate will not suffer

Table 2.1: Simulation results for Cox models with $s_n = 20$, $\alpha_{0j} = 0.35$, and $\rho = 0.5$ under exponential censoring

			PSIS		PSIS-L		PSIS-A		PSIS-S	
% censoring	q_n	$\lvert\hat{\mathcal{M}}\rvert$	FN	FP	FN	FP	FN	FP	FN	FP
20	1e-6	0.86	0.96	3e-7	0.96	0.00	0.96	0.00	0.96	0.00
20	1e-5	2.45	0.89	1e-5	0.89	0.00	0.89	0.00	0.90	0.00
20	1e-4	7.29	0.74	1e-4	0.75	0.00	0.75	0.00	0.78	0.00
20	1e-3	30.74	0.52	1e-3	0.56	0.00	0.57	0.00	0.80	0.00
20	0.01	223.46	0.27	0.01	0.50	0.00	0.51	0.00	0.60	0.00
20	0.10	2066.62	0.07	0.10	0.45	0.00	0.46	0.00	0.64	0.00
20	0.20	4084.34	0.04	0.20	0.45	0.00	0.46	0.00	1.00	0.00
20	0.30	6085.28	0.03	0.30	0.45	0.00	0.46	0.00	1.00	0.00
20	0.40	8087.81	0.02	0.40	0.45	0.00	0.46	0.00	1.00	0.00
20	0.50	10076.05	0.01	0.50	0.45	0.00	0.46	0.00	1.00	0.00
20	0.60	12063.53	0.01	0.60	0.45	0.00	0.46	0.00	1.00	0.00
20	0.70	14049.73	0.01	0.70	0.45	0.00	0.46	0.00	1.00	0.00
20	0.80	16035.60	0.00	0.80	0.45	0.00	0.46	0.00	1.00	0.00
20	0.90	18018.92	0.00	0.90	0.45	0.00	0.46	0.00	1.00	0.00
20	1.00	19999.97	0.00	1.00	0.45	0.00	0.46	0.00	1.00	0.00
50	1e-6	0.46	0.98	8e-7	0.98	0.00	0.98	0.00	0.98	0.00
50	1e-5	1.48	0.93	9e-6	0.93	0.00	0.93	0.00	0.94	0.00
50	1e-4	5.53	0.82	1e-4	0.83	0.00	0.83	0.00	0.87	0.00
50	1e-3	27.48	0.64	1e-3	0.69	0.00	0.70	0.00	0.84	0.00
50	0.01	218.90	0.37	0.01	0.70	0.00	0.71	0.00	0.80	0.00
50	0.10	2055.34	0.11	0.10	0.84	0.00	0.84	0.00	0.82	0.00
50	0.20	4066.57	0.06	0.20	0.85	0.00	0.86	0.00	1.00	0.00
50	0.30	6072.90	0.04	0.30	0.86	0.00	0.87	0.00	1.00	0.00
50	0.40	8071.28	0.03	0.40	0.86	0.00	0.87	0.00	1.00	0.00
50	0.50	10061.46	0.02	0.50	0.87	0.00	0.87	0.00	1.00	0.00
50	0.60	12055.20	0.02	0.60	0.87	0.00	0.87	0.00	1.00	0.00
50	0.70	14048.17	0.01	0.70	0.87	0.00	0.87	0.00	1.00	0.00
50	0.80	16034.58	0.00	0.80	0.87	0.00	0.87	0.00	1.00	0.00
50	0.90	18018.47	0.00	0.90	0.87	0.00	0.87	0.00	1.00	0.00
50	1.00	19999.97	0.00	1.00	0.87	0.00	0.87	0.00	1.00	0.00
70	1e-6	0.04	1.00	1e-6	1.00	0.00	1.00	0.00	1.00	0.00
70	1e-5	0.33	0.99	7e-6	0.99	0.00	0.99	0.00	0.99	0.00
70	1e-4	2.33	0.96	8e-5	0.96	0.00	0.96	0.00	0.97	0.00
70	1e-3	20.34	0.86	9e-4	0.89	0.00	0.89	0.00	0.95	0.00
70	0.01	200.62	0.63	0.01	0.98	0.00	0.98	0.00	0.97	0.00
70	0.10	2036.15	0.28	0.10	0.99	0.00	0.99	0.00	1.00	0.00
70	0.20	4056.93	0.17	0.20	0.99	0.00	0.99	0.00	1.00	0.00
70	0.30	6073.71	0.11	0.30	0.99	0.00	0.99	0.00	1.00	0.00
70	0.40	8074.10	0.08	0.40	0.99	0.00	0.99	0.00	1.00	0.00
70	0.50	10069.87	0.06	0.50	0.99	0.00	0.99	0.00	1.00	0.00
70	0.60	12061.20	0.04	0.60	0.99	0.00	0.99	0.00	1.00	0.00
70	0.70	14053.00	0.03	0.70	0.99	0.00	0.99	0.00	1.00	0.00
70	0.80	16036.78	0.02	0.80	0.99	0.00	0.99	0.00	1.00	0.00
70	0.90	18018.22	0.01	0.90	0.99	0.00	0.99	0.00	1.00	0.00
70	1.00	19999.98	0.00	1.00	0.99	0.00	0.99	0.00	1.00	0.00

Table 2.2: Simulation results for Cox models with $s_n = 20$, $\alpha_{0j} = 0.35$, and $\rho = 0.9$ under exponential censoring

			PSIS		PSIS-L		PSIS-A		PSIS-S	
% censoring	q_n	$\lvert\hat{\mathcal{M}}\rvert$	FN	FP	FN	FP	FN	FP	FN	FP
20	1e-6	20.57	0.03	6e-5	0.26	0.00	0.27	0.00	0.65	0.00
20	1e-5	22.00	0.01	1e-4	0.25	0.00	0.26	0.00	0.64	0.00
20	1e-4	25.39	0.00	3e-4	0.26	0.00	0.27	0.00	0.65	0.00
20	1e-3	46.84	0.00	1e-3	0.29	0.00	0.32	0.00	0.68	0.00
20	0.01	237.31	0.00	0.01	0.43	0.00	0.45	0.00	0.65	0.00
20	0.10	2069.38	0.00	0.10	0.39	0.00	0.41	0.00	0.69	0.00
20	0.20	4082.82	0.00	0.20	0.37	0.00	0.40	0.00	1.00	0.00
20	0.30	6085.60	0.00	0.30	0.38	0.00	0.40	0.00	1.00	0.00
20	0.40	8080.28	0.00	0.40	0.38	0.00	0.40	0.00	1.00	0.00
20	0.50	10072.06	0.00	0.50	0.38	0.00	0.40	0.00	1.00	0.00
20	0.60	12063.09	0.00	0.60	0.38	0.00	0.39	0.00	1.00	0.00
20	0.70	14052.35	0.00	0.70	0.38	0.00	0.40	0.00	1.00	0.00
20	0.80	16041.30	0.00	0.80	0.38	0.00	0.39	0.00	1.00	0.00
20	0.90	18017.65	0.00	0.90	0.38	0.00	0.39	0.00	1.00	0.00
20	1.00	19999.98	0.00	1.00	0.38	0.00	0.40	0.00	1.00	0.00
50	1e-6	19.38	0.07	4e-5	0.35	0.00	0.37	0.00	0.72	0.00
50	1e-5	21.08	0.03	8e-5	0.34	0.00	0.36	0.00	0.71	0.00
50	1e-4	24.35	0.01	2e-4	0.34	0.00	0.36	0.00	0.72	0.00
50	1e-3	44.27	0.00	1e-3	0.30	0.00	0.32	0.00	0.74	0.00
50	0.01	229.40	0.00	0.01	0.50	0.00	0.53	0.00	0.69	0.00
50	0.10	2059.74	0.00	0.10	0.51	0.00	0.54	0.00	0.78	0.00
50	0.20	4075.78	0.00	0.20	0.51	0.00	0.55	0.00	1.00	0.00
50	0.30	6082.62	0.00	0.30	0.51	0.00	0.55	0.00	1.00	0.00
50	0.40	8076.19	0.00	0.40	0.51	0.00	0.54	0.00	1.00	0.00
50	0.50	10069.25	0.00	0.50	0.51	0.00	0.54	0.00	1.00	0.00
50	0.60	12056.07	0.00	0.60	0.51	0.00	0.54	0.00	1.00	0.00
50	0.70	14049.39	0.00	0.70	0.51	0.00	0.55	0.00	1.00	0.00
50	0.80	16035.47	0.00	0.80	0.52	0.00	0.55	0.00	1.00	0.00
50	0.90	18016.56	0.00	0.90	0.52	0.00	0.55	0.00	1.00	0.00
50	1.00	19999.99	0.00	1.00	0.52	0.00	0.55	0.00	1.00	0.00
70	1e-6	17.07	0.17	2e-5	0.49	0.00	0.51	0.00	0.78	0.00
70	1e-5	19.29	0.09	6e-5	0.45	0.00	0.47	0.00	0.77	0.00
70	1e-4	22.88	0.04	2e-4	0.40	0.00	0.43	0.00	0.76	0.00
70	1e-3	42.52	0.01	1e-3	0.35	0.00	0.39	0.00	0.74	0.00
70	1e-2	225.35	0.00	0.01	0.63	0.00	0.66	0.00	0.78	0.00
70	0.10	2047.36	0.00	0.10	0.65	0.00	0.68	0.00	0.86	0.00
70	0.20	4061.41	0.00	0.20	0.65	0.00	0.68	0.00	1.00	0.00
70	0.30	6066.76	0.00	0.30	0.65	0.00	0.68	0.00	1.00	0.00
70	0.40	8069.35	0.00	0.40	0.65	0.00	0.68	0.00	1.00	0.00
70	0.50	10063.83	0.00	0.50	0.65	0.00	0.68	0.00	1.00	0.00
70	0.60	12052.45	0.00	0.60	0.65	0.00	0.68	0.00	1.00	0.00
70	0.70	14041.27	0.00	0.70	0.65	0.00	0.68	0.00	1.00	0.00
70	0.80	16029.24	0.00	0.80	0.65	0.00	0.68	0.00	1.00	0.00
70	0.90	18015.47	0.00	0.90	0.65	0.00	0.68	0.00	1.00	0.00
70	1.00	19999.99	0.00	1.00	0.65	0.00	0.68	0.00	1.00	0.00

Figure 2.1: False negative rates for Cox models with $\alpha_{0j} = 0.35$ (dashes) and $\alpha_{0j} = 0.7$ (solid) under exponential censoring.

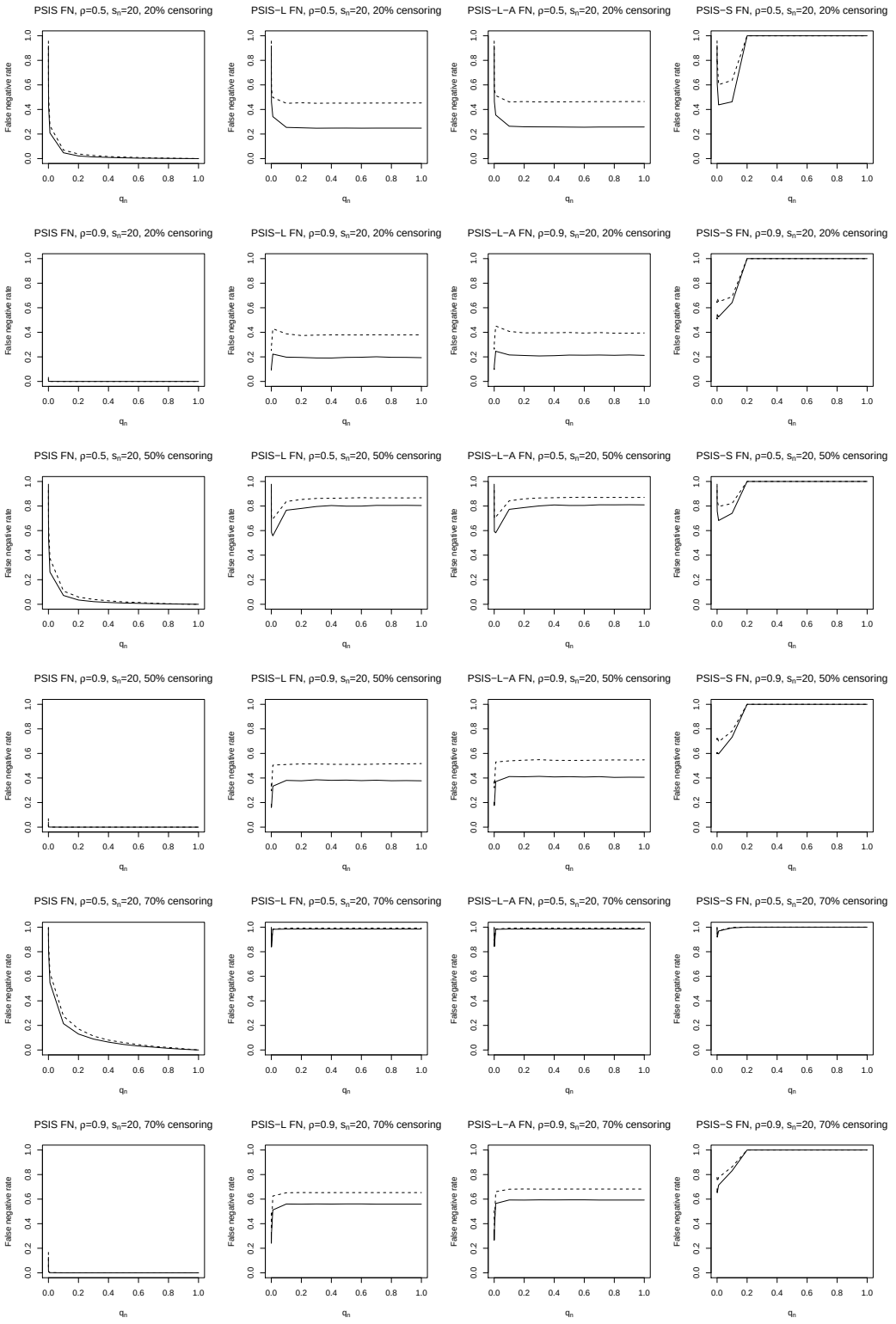

Figure 2.1 (Continued).

much, as long as the amount of censoring is not too high.

The average false negative rates for PSIS-L, PSIS-L-A, and PSIS-S are also plotted in Figure 2.1. The corresponding false positive rates are all very low and so are not plotted (see Tables 2.1 and 2.2). The PSIS-S results give false negative rates of 1 for large q_n because the SCAD algorithm fails when the number of covariates is too large, so we already see the usefulness of PSIS in facilitating computations. For all methods, higher values of α_{0j} exhibit lower false negative rates, and as before we see better performance when the covariates are more correlated.

Interestingly, with 50% censoring, when q_n is small the PSIS-L, PSIS-L-A, and PSIS-S false negative rates are noticeably lower than those after running lasso, lasso-adaptive lasso, or SCAD alone (i.e. $q_n = 1$). This supports the use of PSIS prior to running regularized regression. However, when the censoring rate is fairly low (20%) or fairly high (70%), this effect diminishes, to the point where PSIS actually degrades the performance of the regularized regressions when $\rho = 0.5$ at 20% censoring. This could be because with low censoring, there might already be sufficient data for the regularized regressions to select from the $p_n = 20000$ covariates, even in the absence of PSIS. At the other extreme, when there is 70% censoring, there might be so little data (with $n = 100$) that no regularized method, with or without PSIS, would perform well.

To assess the robustness of our procedure, we also generated 200 datasets from log-normal models. Each dataset had $n = 100$ and $p_n = 20000$, and covariates $\mathbf{Z}_i$ were generated using the same procedure as above, for $\rho = 0.5$ or 0.9. Survival times T_i were generated according to $\log(T_i) = \boldsymbol{\alpha}_0^T \mathbf{Z}_i + \epsilon_i$, where ϵ_i followed a standard normal distribution and $\boldsymbol{\alpha}_0$ had $s_n = 10$ or 20 nonzero elements all equal to either 0.35 or 0.7. Censoring times were generated using the same schemes and rates as before.

Tables 2.3 and 2.4 and Figure 2.2 show that PSIS can still perform very well when the Cox model is misspecified. Again, the numerical results when $s_n = 10$, $\alpha_{0j} = 0.7$, or the

Table 2.3: Simulation results for log-normal models with $s_n = 20$, $\alpha_{0j} = 0.35$, and $\rho = 0.5$ under exponential censoring

			PSIS		PSIS-L		PSIS-A		PSIS-S	
% censoring	q_n	$\lvert\hat{\mathcal{M}}\rvert$	FN	FP	FN	FP	FN	FP	FN	FP
20	1e-6	1.07	0.95	5e-7	0.95	0.00	0.95	0.00	0.95	0.00
20	1e-5	2.92	0.86	1e-5	0.87	0.00	0.87	0.00	0.88	0.00
20	1e-4	7.83	0.71	1e-4	0.72	0.00	0.72	0.00	0.76	0.00
20	1e-3	31.64	0.50	1e-3	0.54	0.00	0.54	0.00	0.74	0.00
20	0.01	225.09	0.25	0.01	0.45	0.00	0.46	0.00	0.54	0.00
20	0.10	2072.03	0.07	0.10	0.38	0.00	0.39	0.00	0.58	0.00
20	0.20	4086.02	0.04	0.20	0.38	0.00	0.39	0.00	1.00	0.00
20	0.30	6091.88	0.02	0.30	0.38	0.00	0.39	0.00	1.00	0.00
20	0.40	8096.85	0.01	0.40	0.38	0.00	0.39	0.00	1.00	0.00
20	0.50	10084.36	0.01	0.50	0.38	0.00	0.39	0.00	1.00	0.00
20	0.60	12071.20	0.01	0.60	0.38	0.00	0.39	0.00	1.00	0.00
20	0.70	14053.60	0.00	0.70	0.38	0.00	0.39	0.00	1.00	0.00
20	0.80	16035.81	0.00	0.80	0.38	0.00	0.39	0.00	1.00	0.00
20	0.90	18014.10	0.00	0.90	0.38	0.00	0.39	0.00	1.00	0.00
20	1.00	19999.96	0.00	1.00	0.38	0.00	0.39	0.00	1.00	0.00
50	1e-6	0.58	0.97	2e-6	0.97	0.00	0.97	0.00	0.97	0.00
50	1e-5	1.68	0.92	8e-6	0.92	0.00	0.92	0.00	0.93	0.00
50	1e-4	5.74	0.80	9e-5	0.81	0.00	0.81	0.00	0.85	0.00
50	1e-3	28.12	0.61	1e-3	0.66	0.00	0.66	0.00	0.82	0.00
50	0.01	217.45	0.35	0.01	0.63	0.00	0.65	0.00	0.74	0.00
50	0.10	2053.91	0.10	0.10	0.79	0.00	0.80	0.00	0.79	0.00
50	0.20	4066.55	0.05	0.20	0.81	0.00	0.82	0.00	1.00	0.00
50	0.30	6071.31	0.03	0.30	0.81	0.00	0.82	0.00	1.00	0.00
50	0.40	8071.39	0.02	0.40	0.82	0.00	0.82	0.00	1.00	0.00
50	0.50	10067.45	0.01	0.50	0.82	0.00	0.83	0.00	1.00	0.00
50	0.60	12061.88	0.01	0.60	0.82	0.00	0.83	0.00	1.00	0.00
50	0.70	14045.89	0.00	0.70	0.82	0.00	0.83	0.00	1.00	0.00
50	0.80	16030.90	0.00	0.80	0.82	0.00	0.83	0.00	1.00	0.00
50	0.90	18017.08	0.00	0.90	0.82	0.00	0.83	0.00	1.00	0.00
50	1.00	19999.99	0.00	1.00	0.82	0.00	0.83	0.00	1.00	0.00
70	1e-6	0.07	1.00	8e-7	1.00	0.00	1.00	0.00	1.00	0.00
70	1e-5	0.45	0.98	7e-6	0.98	0.00	0.98	0.00	0.99	0.00
70	1e-4	2.62	0.95	8e-5	0.95	0.00	0.95	0.00	0.96	0.00
70	1e-3	20.59	0.83	9e-4	0.86	0.00	0.87	0.00	0.93	0.00
70	0.01	201.80	0.59	0.01	0.98	0.00	0.98	0.00	0.96	0.00
70	0.10	2032.38	0.24	0.10	0.99	0.00	0.99	0.00	1.00	0.00
70	0.20	4057.72	0.15	0.20	0.99	0.00	0.99	0.00	1.00	0.00
70	0.30	6071.12	0.10	0.30	0.99	0.00	0.99	0.00	1.00	0.00
70	0.40	8075.48	0.07	0.40	0.99	0.00	0.99	0.00	1.00	0.00
70	0.50	10069.12	0.05	0.50	0.99	0.00	0.99	0.00	1.00	0.00
70	0.60	12056.30	0.04	0.60	0.99	0.00	0.99	0.00	1.00	0.00
70	0.70	14044.67	0.03	0.70	0.99	0.00	0.99	0.00	1.00	0.00
70	0.80	16029.99	0.02	0.80	0.99	0.00	0.99	0.00	1.00	0.00
70	0.90	18015.69	0.01	0.90	0.99	0.00	0.99	0.00	1.00	0.00
70	1.00	19999.99	0.00	1.00	0.99	0.00	0.99	0.00	1.00	0.00

Table 2.4: Simulation results for log-normal models with $s_n = 20$, $\alpha_{0j} = 0.35$, and $\rho = 0.9$ under exponential censoring

			PSIS		PSIS-L		PSIS-A		PSIS-S	
% censoring	q_n	$\lvert\hat{\mathcal{M}}\rvert$	FN	FP	FN	FP	FN	FP	FN	FP
20	1e-6	20.63	0.03	6e-5	0.25	0.00	0.25	0.00	0.63	0.00
20	1e-5	21.98	0.01	1e-4	0.23	0.00	0.24	0.00	0.62	0.00
20	1e-4	24.94	0.00	2e-4	0.23	0.00	0.25	0.00	0.63	0.00
20	1e-3	45.60	0.00	1e-3	0.26	0.00	0.29	0.00	0.65	0.00
20	0.01	232.98	0.00	0.01	0.35	0.00	0.37	0.00	0.54	0.00
20	0.10	2064.58	0.00	0.10	0.31	0.00	0.33	0.00	0.64	0.00
20	0.20	4079.41	0.00	0.20	0.30	0.00	0.33	0.00	1.00	0.00
20	0.30	6089.45	0.00	0.30	0.30	0.00	0.32	0.00	1.00	0.00
20	0.40	8089.52	0.00	0.40	0.30	0.00	0.32	0.00	1.00	0.00
20	0.50	10075.52	0.00	0.50	0.30	0.00	0.33	0.00	1.00	0.00
20	0.60	12064.03	0.00	0.60	0.30	0.00	0.32	0.00	1.00	0.00
20	0.70	14048.61	0.00	0.70	0.30	0.00	0.32	0.00	1.00	0.00
20	0.80	16031.75	0.00	0.80	0.30	0.00	0.33	0.00	1.00	0.00
20	0.90	18014.60	0.00	0.90	0.31	0.00	0.33	0.00	1.00	0.00
20	1.00	19999.98	0.00	1.00	0.31	0.00	0.33	0.00	1.00	0.00
50	1e-6	19.65	0.06	4e-5	0.32	0.00	0.34	0.00	0.68	0.00
50	1e-5	21.14	0.02	8e-5	0.30	0.00	0.32	0.00	0.68	0.00
50	1e-4	24.32	0.01	2e-4	0.30	0.00	0.32	0.00	0.68	0.00
50	1e-3	44.18	0.00	1e-3	0.24	0.00	0.25	0.00	0.69	0.00
50	0.01	229.64	0.00	0.01	0.43	0.00	0.46	0.00	0.63	0.00
50	0.10	2056.83	0.00	0.10	0.46	0.00	0.50	0.00	0.72	0.00
50	0.20	4067.93	0.00	0.20	0.47	0.00	0.50	0.00	1.00	0.00
50	0.30	6075.55	0.00	0.30	0.47	0.00	0.50	0.00	1.00	0.00
50	0.40	8067.98	0.00	0.40	0.47	0.00	0.50	0.00	1.00	0.00
50	0.50	10065.84	0.00	0.50	0.47	0.00	0.51	0.00	1.00	0.00
50	0.60	12059.20	0.00	0.60	0.46	0.00	0.50	0.00	1.00	0.00
50	0.70	14048.42	0.00	0.70	0.46	0.00	0.50	0.00	1.00	0.00
50	0.80	16028.41	0.00	0.80	0.47	0.00	0.50	0.00	1.00	0.00
50	0.90	18014.56	0.00	0.90	0.47	0.00	0.50	0.00	1.00	0.00
50	1.00	19999.97	0.00	1.00	0.47	0.00	0.50	0.00	1.00	0.00
70	1e-6	17.46	0.14	2e-5	0.41	0.00	0.44	0.00	0.75	0.00
70	1e-5	19.54	0.07	5e-5	0.37	0.00	0.39	0.00	0.73	0.00
70	1e-4	22.92	0.03	2e-4	0.32	0.00	0.35	0.00	0.72	0.00
70	1e-3	43.28	0.01	1e-3	0.35	0.00	0.38	0.00	0.69	0.00
70	0.01	229.22	0.00	0.01	0.57	0.00	0.60	0.00	0.75	0.00
70	0.10	2062.75	0.00	0.10	0.63	0.00	0.65	0.00	0.86	0.00
70	0.20	4080.23	0.00	0.20	0.63	0.00	0.65	0.00	1.00	0.00
70	0.30	6083.23	0.00	0.30	0.62	0.00	0.65	0.00	1.00	0.00
70	0.40	8080.86	0.00	0.40	0.62	0.00	0.65	0.00	1.00	0.00
70	0.50	10073.59	0.00	0.50	0.62	0.00	0.65	0.00	1.00	0.00
70	0.60	12067.05	0.00	0.60	0.62	0.00	0.65	0.00	1.00	0.00
70	0.70	14053.72	0.00	0.70	0.62	0.00	0.65	0.00	1.00	0.00
70	0.80	16038.41	0.00	0.80	0.62	0.00	0.65	0.00	1.00	0.00
70	0.90	18019.44	0.00	0.90	0.62	0.00	0.65	0.00	1.00	0.00
70	1.00	19999.99	0.00	1.00	0.62	0.00	0.65	0.00	1.00	0.00

Figure 2.2: False negative rates for log-normal models with $s_n = 20$, $\alpha_{0j} = 0.35$ (dashes), and $\alpha_{0j} = 0.7$ (solid) under exponential censoring.

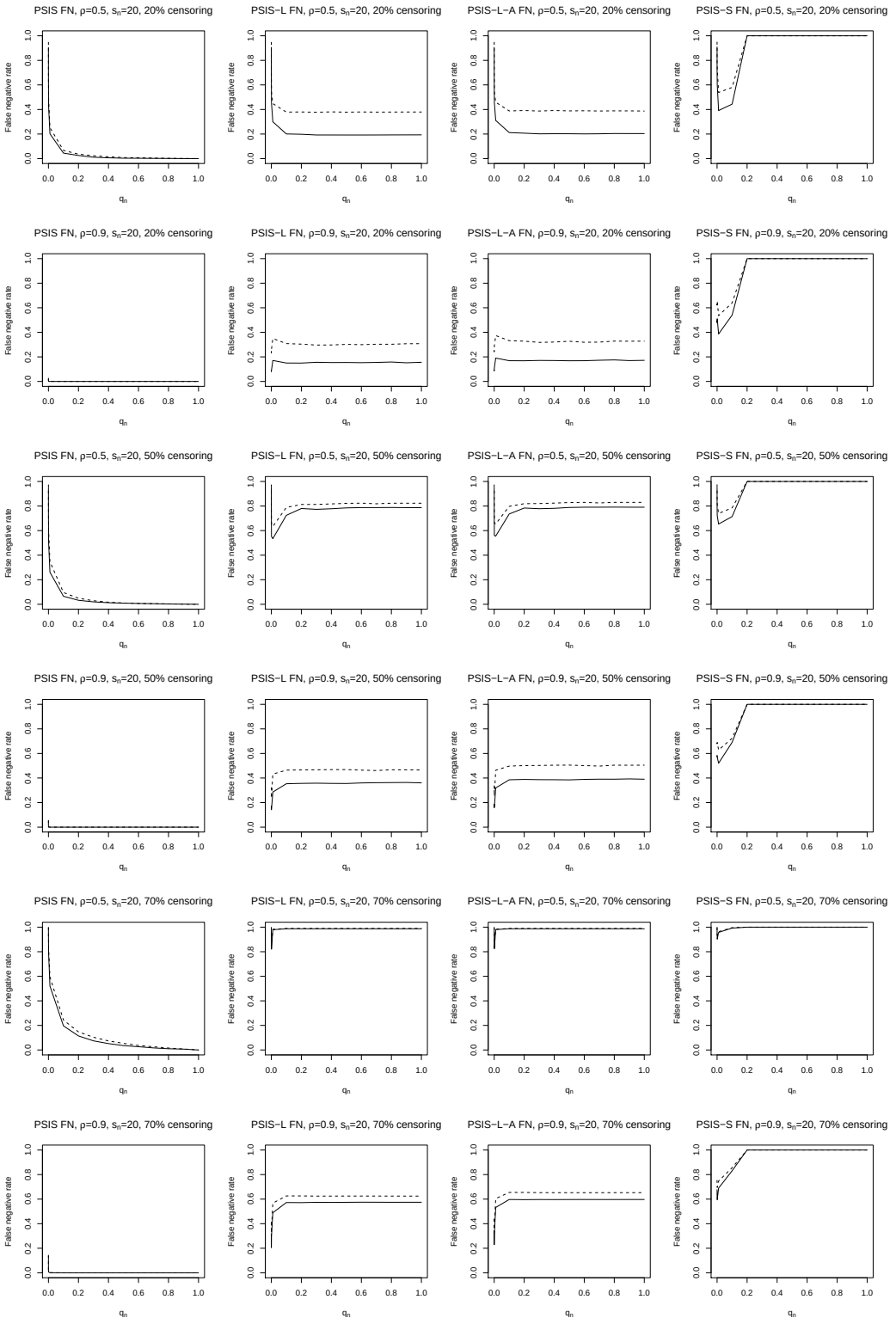

Figure 2.2 (Continued).

censoring times were uniformly distributed are omitted from the tables to save space. These results follow the same trends as those of the correctly specified simulations discussed above. In particular, the principled cutoff procedure still shows good performance, and using PSIS when q_n is small can still lead to lower false negative rates than when $q_n = 1$.

2.5 Analysis of the myeloma study

Recent advances in understanding the biological mechanisms underlying multiple myeloma have offered new possibilities for therapy (Hideshima et al., 2007). Time-to-event outcomes offer information about the progression of the disease, and in this vein several studies have examined the relationship between gene expression levels and survival (Decaux et al., 2008). In one such study conducted by Millennium Pharmaceuticals (Mulligan et al., 2007), mRNA expression levels were collected using Affymetrix U133A/B arrays from myeloma cells of 80 patients enrolled in a clinical trial of bortezomib (accession number GSE9782, trial 39). Median survival time was 684 days after randomization, and 50% of the observations were censored. We applied our methods to this data.

Expression values were measured for 44760 probesets, encompassing more than 22000 genes, and were $\log_2$-transformed. We performed PSIS and chose $q_n = 1/44760$, for two reasons. First, our simulation results suggest that for large ρ, the true false positive rate can be larger than our nominal level, and genetic datasets are probably highly correlated. Second, gene expression levels are very likely related to the survival outcomes, but only weakly so. Many of our genes are probably not sufficiently important, in the sense of Assumption 11, so allowing even a small false positive rate would result in including a huge number of genes. For these reasons, we want to control the false positive rate to the extent possible, but on the other hand we cannot allow $f = 0$ or else $\gamma_n = \infty$. Thus we considered $f = 1$, which leads to our choice of q_n.

Table 2.5: Predictive accuracies using myeloma data

	PSIS, $q_n = 0.0001$		Random probesets		All probesets	
Method	C-stat (SD)	Size (SD)	C-stat (SD)	Size (SD)	C-stat (SD)	Size (SD)
Lasso	0.60 (0.09)	11.43 (9.18)	0.15 (0.25)	3.33 (10.74)	0.33 (0.33)	20.91 (24.07)
Lasso-alasso	0.60 (0.09)	11.13 (8.97)	0.15 (0.25)	3.29 (10.60)	0.33 (0.33)	18.86 (21.74)
SCAD	0.60 (0.09)	7.13 (6.67)	0.33 (0.28)	2.48 (7.43)	—	—

We could not directly evaluate the performance of PSIS without knowing which genes are "truly" important. Instead, we ran PSIS to get a screened model $\hat{\mathcal{M}}$ and also randomly selected $|\hat{\mathcal{M}}|$ probesets. We then compared the prediction accuracies of PSIS-L, PSIS-L-A, and PSIS-S to those obtained by fitting lasso, lasso-adaptive lasso, and SCAD on the randomly selected probesets. Using random genes as negative controls is common in these type of experiments (Hofmann et al., 2002; Aerts et al., 2006; Fan et al., 2010). We also fit the regularized regression procedures on the full dataset, without any screening (i.e. $q_n = 1$).

For each of these methods, we randomly partitioned the data into a 60-patient training set and a 20-patient testing set. We then used the models fit in the training set to calculate scores for each subject in the testing set, and evaluated the predictive performance using the C-statistic (Uno et al., 2011b). We repeated this entire process 200 times. Better performances from the screened methods would provide evidence that PSIS is indeed finding predictively important genes.

Table 2.5 reports the average C-statistics and model sizes obtained by our different methods. We see that PSIS-L, PSIS-L-A, and PSIS-S perform much better than the corresponding regressions fit using randomly selected probesets. When we do not screen the data, SCAD fails, and lasso and lasso-adaptive lasso do not perform as well as the screened versions.

Because of the selection consistency of the adaptive lasso (Zou, 2006; Zhang and Lu, 2007), we next applied PSIS-L-A with $q_n = 1/44760$ to all 80 patients. Table 2.6 gives the probesets we found to have nonzero parameter estimates, as well as their estimated coefficients. Indeed, our results include some genes previously found to be related to myeloma, e.g. IGHV3-23 (Hadzidimitriou et al., 2006) and PRKDC (Shaughnessy and Barlogie, 2003). Fi-

Table 2.6: Genes found using PSIS-L-A, $q_n = 0.0001$

Probeset	Gene	Description	Coefficient
219999_at	MAN2A2	mannosidase, alpha, class 2A, member 2	-0.27
207677_s_at	NCF4	neutrophil cytosolic factor 4, 40kDa	-0.14
216510_x_at	IGHV3-23	immunoglobulin heavy variable 3-23	-0.35
222610_s_at	S100PBP	S100P binding protein	0.15
203550_s_at	FAM189B	family with sequence similarity 189, member B	0.06
208694_at	PRKDC	protein kinase, DNA-activated, catalytic polypeptide	0.12
223277_at	C3orf75	chromosome 3 open reading frame 75	0.18
234980_at	TMEM56	transmembrane protein 56	-0.37
213893_x_at	PMS2L5	postmeiotic segregation increased 2-like 5	0.26
217518_at	MYOF	myoferlin	-0.13
202587_s_at	AK1	adenylate kinase 1	-0.11
232452_at	LOC148824	hypothetical LOC148824	0.42
209217_s_at	WDR45	WD repeat domain 45	-1.29
226692_at	SERF2	small EDRK-rich factor 2	-1.15
223114_at	COQ5	coenzyme Q5 homolog, methyltransferase	0.29

nally, we evaluated the predictive performance of this model using an independent validation dataset (accession number GSE9782, trials 24, 25, and 40). The model in Table 2.6 achieved a C-statistic of 0.59, which matches the cross-validation estimate of Table 2.5. These results indicate that PSIS is an effective way to identify predictively important genes while controlling the false positive rate, and that implementing PSIS before regularized regression can lead to more computationally amenable, interpretable models with high predictive power.

2.6 Discussion

This paper advances the field in three distinct ways. First, we have demonstrated that with censored outcomes, sure independence screening using marginal Cox regressions is a theoretically justified, effective way to reduce ultra-high-dimensional data to moderate sizes before applying more sophisticated variable selection procedures. In particular, we have described new, necessary condition on the dependence between the covariates and the censoring distribution. Second, we have provided a simple, principled method to select the number of variables to retain after screening and illustrated its effectiveness with simulated data. Our procedure could be easily extended to other screening methods. Finally, we have demonstrated through the motivating myeloma example that pre-screening may improve risk classification and identify predictive genes. There are a number of ways to broaden the scope

of our method. So far we have dealt only with covariates that are constant in time, and we have not considered tied observations. Our method could also be extended to multivariate survival, competing risks, and other extensions of the Cox model.

While our simulations suggest that PSIS performs well even with correlated covariates, it would be interesting to explore other screening methods proposed specifically to deal with this situation. One approach is the ISIS method of Fan and Lv (2008), which starts with an initial model of potentially important covariates, regresses the residuals from the working model on each of the remaining covariates to expand the working model, and iterates this process in order to capture any important covariates that would be missed in univariate screening. Residuals are unavailable with censored observations, but Fan et al. (2010) generalized this iterative idea by working instead with log-likelihood ratios. Their formulation is easily applied to the log-partial likelihood of the Cox model, which they have implemented in the R package `SIS`. However, the theoretical properties of this procedure have not been investigated.

Finally, our theoretical analysis of sure independence screening touches on some philosophical questions about notions of variable importance. Biological phenomena often arise from the complex interactions of genes and other factors whose individual effects can be fairly weak but still non-zero. Thus merely having a non-zero contribution to the model is not a useful notion of importance, because then nearly every variable would be important. It may be more useful to conceive of importance as a finite sample property, in the sense that covariates whose signals are higher than the noise level of the estimator being used are to be considered important. In our method, for instance, the so-called important covariates satisfy Theorem 4, or else they could not be detected by marginal Cox regressions. Perhaps a good variable selection technique is one that, instead of selecting every variable with a non-zero contribution to the outcome, retains only those variables that, for a given n, meet the finite-sample definition of importance as defined in Theorem 4. The sure screening property of our method indicates that as n increases, we get closer to achieving this goal.

Acknowledgements

We thank Professor Jianqing Fan for reading an earlier version of this article and for many helpful suggestions that substantially improved the manuscript.

2.7 Appendix A: Assumptions

Let the true hazard function $\lambda(x; \mathbf{Z}_i)$ be given by (2.1), and denote the true survival functions of T_i and C_i as $S_T(x; \mathbf{Z}_i) = \exp\{-\exp(\alpha_0^T \mathbf{Z}_i)\Lambda_0(x)\}$ and $S_C(x; \mathbf{Z}_i) = \mathrm{P}(C_i > x|\mathbf{Z}_i)$, where the cumulative baseline hazard function $\Lambda_0(x) = \int_0^x \lambda_0(s)ds$. To conserve space we will write these as S_T and S_C. For simplicity we will drop the subject-specific subscripts i, except in the proof of Theorem 3. We will also need the following assumptions. We use notation introduced in Section 2.3.1.

Assumption 5 *There exists a neighborhood B of β_{0j} such that for each $t < \infty$,*

$$\sup_{x\in[0,t],\beta\in B} |S_j^{(0)}(\beta, x) - s_j^{(0)}(\beta, x)| \to 0 \tag{2.10}$$

in probability as $n \to \infty$, $s_j^{(0)}(\beta, x)$ is bounded away from zero on $B \times [0, t]$, and $s_j^{(0)}(\beta, x)$ and $s_j^{(1)}(\beta, x)$ are bounded on $B \times [0, t]$.

Assumption 6 *For each $t < \infty$ and $j = 1, \ldots, p_n$, $\int_0^t s_j^{(2)}(x)dx < \infty$.*

Assumption 7 *The true parameter vector $\boldsymbol{\alpha}_0$ belongs to a compact set such that each component α_{0j} is bounded by a constant $A > 0$. Furthermore, $\|\boldsymbol{\alpha}_0\|_1$ is bounded by a constant $L > 0$.*

Assumption 8 *With τ (the study duration) as defined in Section 2.3.1, $\Lambda_0(\tau)$ is bounded by a positive constant.*

Assumption 9 *There is some constant $H > 0$ such that $n^{-1}|U_j(\hat{\beta}_j) - U_j(\beta_{0j})| \geq H|\hat{\beta}_j - \beta_{0j}|$ for all $j = 1, \ldots, p_n$.*

Assumptions 5 and 6 are standard in survival analysis. Assumption 7 controls the total effect size of the covariates, which intuitively should be bounded and independent of sample size. The bounded cumulative baseline hazard function required by Assumption 8 usually holds in practice. Finally, Assumption 9 is reasonable because by the mean value theorem, we know that $n^{-1}|U_j(\hat{\beta}_j) - U_j(\beta_{0j})| = |n^{-1}I_j(\beta^*)||\hat{\beta}_j - \beta_{0j}|$ for some β^* between $\hat{\beta}_j$ and β_{0j}. It can be shown that $I_j(\beta^*)$ converges to the absolute value of the limiting information $-\partial u_j(\beta)/\partial\beta$ evaluated at the true β_{0j} (Fleming and Harrington, 2005), and it is reasonable to assume that this limiting information is bounded from below away from zero. Thus for n sufficiently large, we can take $H = \inf_{\beta,j} |\partial u_j(\beta)/\partial\beta|$ such that $H \neq 0$.

Our PSIS method will have good theoretical properties if the covariates $\mathbf{Z}_i$ also satistify the following reasonable assumptions. Versions of these assumptions have been previously proposed (Fan and Lv, 2008; Fan and Song, 2010), but modifications are required when working with censored data.

Assumption 10 *The Z_{ij} are independent of time and bounded by a constant $K > 0$, and $\mathrm{E}(Z_{ij}) = 0$ for all j.*

Assumption 11 *If $F_T(x; \mathbf{Z}_i)$ is the cumulative distribution function of T_i given $\mathbf{Z}_i$, then for constants $c_1 > 0$ and $\kappa < 1/2$, $\min_{j\in\mathcal{M}} |\mathrm{cov}[Z_{ij}, \mathrm{E}\{F_T(C_i; \mathbf{Z}_i) \mid \mathbf{Z}_i\}]| \geq c_1 n^{-\kappa}$.*

Assumption 12 *The Z_{ij}, $j \in \mathcal{M}^c$ are independent of the Z_{ij}, $j \in \mathcal{M}$ and of C_i.*

The validity of our proposed screening procedure hinges on whether the misspecified marginal Cox regressions can reflect the importance of the corresponding covariates in the joint model. In general it is difficult to directly link the true α_{0j} to the marginal β_{0j} because of the phenomenon of unfaithfulness (Wasserman and Roeder, 2009), where the marginal correlation of Z_{ij} with the outcome can be zero even if α_{0j} is large, due to correlated covariates. Assumption 11 protects against unfaithfulness. Though the outcome is unobservable under censoring, $F_T(C_i; \mathbf{Z}_i)$ is the probability of observing a failure given $\mathbf{Z}_i$ and is a sensible surrogate. Assumption 12 is similar to the partial orthogonality assumption introduced in Fan and Song (2010).

2.8 Appendix B: Proofs

2.8.1 Proof of Theorem 2

We first relate β_{0j} to $\text{cov}[Z_j, \text{E}\{F_T(C; Z) \mid \mathbf{Z}\}]$. Assumptions 11 and 12 will then relate the covariance to α_{0j}.

First, integration by parts gives that

$$\int_0^\tau \text{E}\{Z_j \lambda_0(x) \exp(\boldsymbol{\alpha}_0^T \mathbf{Z}) S_T S_C\} dx = \text{cov}[Z_j, \text{E}\{F_T(C; \mathbf{Z}) \mid \mathbf{Z}\}]. \tag{2.11}$$

Next, we define the function

$$f(\beta) = \int_0^\tau \frac{\text{E}\{Z_j \exp(\beta Z_j) S_T S_C\}}{\text{E}\{\exp(\beta Z_j) S_T S_C\}} \text{E}\{\lambda_0(x) \exp(\boldsymbol{\alpha}_0^T \mathbf{Z}) S_T S_C\} dx. \tag{2.12}$$

Then since β_{0j} is the solution to the estimating equation $u_j(\beta)$ (2.5), we know that $\text{cov}[Z_j, \text{E}\{F_T(C; \mathbf{Z}) \mid \mathbf{Z}\}] = f(\beta_{0j})$. We can use the Cauchy-Schwarz inequality to show that $\partial f(\beta)/\partial \beta \geq 0$, with equality if and only if $\text{P}\{Z_j \exp(\beta Z_j/2)(S_T S_C)^{1/2} = c \exp(\beta Z_j/2)(S_T S_C)^{1/2}\} = 1$ for some constant c. Since this will not hold if Z_j is not constant, we see that $f(\beta)$ is a monotone-increasing function in β.

Now suppose $\alpha_{0j} = 0$ so that $j \in \mathcal{M}^c$. By Assumption 12, Z_j is independent of $\mathrm{E}\{F_T(C;\mathbf{Z}) \mid \mathbf{Z}\}$, so that $f(\beta_{0j}) = \mathrm{cov}[Z_j, \mathrm{E}\{F_T(C;\mathbf{Z}) \mid \mathbf{Z}\}] = 0$. However, we also have that $f(0) = 0$, since $\mathrm{E}\{Z_j S_T S_C\} = \mathrm{E}(Z_j)\mathrm{E}(S_T S_C) = 0$ by Assumption 10 and because Z_j and C are independent for $j \in \mathcal{M}^c$. Because $f(\beta)$ is monotone we know that there is only one value of β such that $f(\beta) = 0$, so that $\beta_{0j} = 0$. Similarly, suppose that $\alpha_{0j} \neq 0$ so that $j \in \mathcal{M}$. Then by Assumption 11, $|f(\beta_{0j})| = |\mathrm{cov}[Z_j, \mathrm{E}\{F_T(C;\mathbf{Z}) \mid \mathbf{Z}\}]| > c_1 n^{-\kappa}$. Therefore $\beta_{0j} \neq 0$ by monotonicity, and we can conclude that $\alpha_{0j} = 0$ if and only if $\beta_{0j} = 0$.

2.8.2 Proof of Theorem 3

We first bound $|U_j(\hat{\beta}_j) - U_j(\beta_{0j})|$ by the supremum of an empirical process, where $U_j(\beta)$ was defined in (2.4). We then use the concentration theorem of Massart (2000) to derive a maximal inequality. We will conclude by using Assumption 9 to extend this inequality to $|\hat{\beta}_j - \beta_{0j}|$.

First, let $\bar{U}_j(\beta) = n^{-1}U_j(\beta)$. Since we still have $\bar{U}_j(\hat{\beta}_j) = 0$, we can write $|\bar{U}_j(\hat{\beta}_j) - \bar{U}_j(\beta_{0j})| = |\bar{U}_j(\beta_{0j})|$. Because $\bar{U}_j(\beta_{0j})$ is not a sum of independent terms, we cannot directly apply empirical process techniques. However, we know from Lin and Wei (1989) that $\bar{U}_j(\beta_{0j}) = n^{-1}\sum_{i=1}^n w_i^{(j)}(\beta_{0j}) + o_p(1)$, where

$$w_i^{(j)}(\beta_{0j}) = \int_0^\tau \left\{ Z_{ij} - \frac{\mathrm{E}\{Z_{ij}\exp(\beta_{0j}Z_{ij})S_T S_C\}}{\mathrm{E}\{\exp(\beta_{0j}Z_{ij})S_T S_C\}} \right\} dN_i(x) - \tag{2.13}$$

$$\int_0^\tau \frac{Y_i(x)\exp(\beta_{0j}Z_{ij})}{\mathrm{E}\{\exp(\beta_{0j}Z_{ij})S_T S_C\}} \left\{ Z_{ij} - \frac{\mathrm{E}\{Z_{ij}\exp(\beta_{0j}Z_{ij})S_T S_C\}}{\mathrm{E}\{\exp(\beta_{0j}Z_i)S_T S_C\}} \right\} E\{dN_i(x)\}. \tag{2.14}$$

and the $w_i^{(j)}(\beta_{0j})$ are independent. Furthermore, it is easy to show that $\mathrm{E}\{w_i^{(j)}(\beta_{0j})\} = 0$. If we let E_n denote the empirical measure, then we can write $|\bar{U}_j(\hat{\beta}_j) - \bar{U}_j(\beta_{0j})| \leq \sup_\beta |(\mathrm{E}_n - \mathrm{E})w_i^{(j)}(\beta)| + o_p(1)$. Thus $|\bar{U}_j(\hat{\beta}_j) - \bar{U}_j(\beta_{0j})|$ is bounded by the sum of the supremum of an empirical process and a term that converges to zero in probability.

To derive a maximal inequality for this process, we first find a bound on $w_i^{(j)}(\beta)$ uniform over β and $j = 1, \ldots, p_n$. Using Assumptions 7, 8, and 10, we can write that

$|w_i^{(j)}(\beta)| \leq 2K[1+\Lambda_0(\tau)\exp\{2K(A+L)\}]$ for $j = 1, \ldots, p_n$. Next, we must find a bound on the expected value of our supremum. Let ε_i, $i = 1, \ldots, n$ be an independent, identically distributed sequence of random variables taking values ± 1 with probability 1/2. In particular, they are independent of $\mathbf{Z}$. Then $\mathrm{E}\{\sup_\beta |(\mathrm{E}_n - \mathrm{E})w_i^{(j)}(\beta)|\} \leq 2\mathrm{E}[\sup_\beta |\mathrm{E}_n\{\varepsilon_i w_i^{(j)}(\beta)\}|]$, by Lemma 2.3.1 of van der Vaart and Wellner (1996). But by the Cauchy-Schwarz inequality, independence of ε_i and Z_i, and the bound on $|w_i^{(j)}(\beta)|$ derived above, we can show that the right side is bounded by $4K[1+\Lambda_0(\tau)\exp\{2K(A+L)\}]\{\mathrm{var}(n^{-1}\sum_{i=1}^n \varepsilon_i)\}^{1/2}$. Then from the concentration theorem of Massart (2000), we know that

$$\begin{aligned} &\mathrm{P}\left[\sup_\beta |(\mathrm{E}_n - \mathrm{E})w_i^{(j)}(\beta)| \geq n^{-1/2}4K[1+\Lambda_0(\tau)\exp\{2K(A+L)\}](1+t)\right] \\ &\leq \exp(-t^2/2). \end{aligned} \tag{2.15}$$

Finally, we can relate this inequality back to $|\hat{\beta}_j - \beta_{0j}|$ with Assumption 9, though we must also deal with the $o_p(1)$ term. Using a previously proven inequality, $|\hat{\beta}_j - \beta_{0j}| \leq H^{-1}\sup_\beta |(\mathrm{E}_n - \mathrm{E})w_i^{(j)}(\beta)| + o_p(1)$, so we can write

$$\begin{aligned} &\mathrm{P}\left[\sqrt{n}|\hat{\beta}_j - \beta_{0j}| \geq 4K[1+\Lambda_0(\tau)\exp\{2K(A+L)\}](1+t)/H\right] \leq \\ &\mathrm{P}\left[\sup_\beta |(\mathrm{E}_n - \mathrm{E})w_i^{(j)}(\beta)| + o_p(1) \geq n^{-1/2}4K[1+\Lambda_0(\tau)\exp\{2K(A+L)\}](1+t)\right]. \end{aligned} \tag{2.16}$$

But for any $\epsilon > 0$, $\mathrm{P}(A+B \geq c) \leq \mathrm{P}(A \geq c-\epsilon) + \mathrm{P}(B \geq \epsilon)$, where A and B are random variables and c is a constant. We conclude by combining this with (2.15) and (2.16) and taking ϵ arbitrarily close to 0.

2.8.3 Proof of Theorem 4

From Theorem 2, we know that $\beta_{0j} \neq 0$ if $j \in \mathcal{M}$. Then by Theorem 2.1 of Struthers and Kalbfleisch (1986) and the mean value theorem, we know that $|u_j(0)| = |u_j(\beta_{0j}) - u_j(0)| = |u_j'(\beta^*)||\beta_{0j}|$ for some β^* between β_{0j} and 0, where $u_j'(\beta) = du_j(\beta)/d\beta$. We will first bound $u_j'(\beta)$ and then use Assumption 11 to conclude.

Integrating by parts, we can show that $|u_j'(\beta)| \leq 2K^2|\mathrm{E}\{S_T(C;\mathbf{Z}) \mid \mathbf{Z}\}|$. But $\mathrm{E}\{S_T(C;\mathbf{Z}) \mid \mathbf{Z}\}$ is bounded by 1, so

$$|\beta_{0j}| \geq 0.5K^{-2}\left|\mathrm{cov}[Z_j, \mathrm{E}\{F_T(C;\mathbf{Z}) \mid \mathbf{Z}\}] - \int_0^\tau \frac{\mathrm{E}(Z_j S_T S_C)}{\mathrm{E}(S_T S_C)}\mathrm{E}\{\lambda_0(x)\exp(\boldsymbol{\alpha}_0^T\mathbf{Z})S_T S_C\}dx\right|. \tag{2.17}$$

Because $S_T S_C$ is the probability of being at risk at time x, we can intuitively see, and also prove, that $\mathrm{E}(Z_j S_T S_C) = \mathrm{cov}(Z_j, S_T S_C)$ and $\mathrm{cov}[Z_j, \mathrm{E}\{F_T(C;\mathbf{Z}) \mid \mathbf{Z}\}]$ have opposite signs. This implied that $j \in \mathcal{M}$, $|\beta_{0j}| \geq 0.5K^{-2}|\mathrm{cov}[Z_j, \mathrm{E}\{F_T(C;\mathbf{Z}) \mid \mathbf{Z}\}]|$, and Assumption 11 gives $\min_{j\in\mathcal{M}}|\beta_{0j}| \geq c_2 n^{-\kappa}$ for $c_2 = 0.5K^{-2}c_1$.

2.8.4 Proof of Theorem 5

We first derive a probability bound for the standardized marginal regression parameters. We can then use this bound to find $\mathrm{P}(\mathcal{M} \subseteq \hat{\mathcal{M}})$.

Let $1 + t = c_2 H n^{1/2-\kappa}/(8K[1 + \Lambda_0(\tau)\exp\{2K(A + L)\}])$, with c_2 and κ as defined in Theorem 4 and $K, \Lambda_0(\tau), A$, and L as defined in Theorem 3. Then by Theorem 3 there exists a constant c_3 such that $\mathrm{P}(|\hat{\beta}_j - \beta_{0j}| \geq c_2 n^{-\kappa}/2) \leq \exp(-c_3 n^{1-2\kappa})$.

If we now set our cutoff $\gamma_n = \Phi^{-1}(1 - q_n/2)$, then we can write the probability of retaining the important covariates as $1 - \mathrm{P}\{\min_{j\in\mathcal{M}} I_j(\hat{\beta}_j)^{1/2}|\hat{\beta}_j| < \gamma_n\} \geq 1 - \mathrm{P}\{\min_{j\in\mathcal{M}}|\hat{\beta}_j| \leq \gamma_n(Hn)^{-1/2}\}$. Using Theorem 4 we can show that $c_2 n^{-\kappa} - |\hat{\beta}_j| \leq |\beta_{0j} - \hat{\beta}_j|$, $j \in \mathcal{M}$, so $\mathrm{P}\{\min_{j\in\mathcal{M}}|\hat{\beta}_j| \leq \gamma_n(Hn)^{-1/2}\} \leq \mathrm{P}\{\max_{j\in\mathcal{M}}|\hat{\beta}_j - \beta_{0j}| \geq c_2 n^{-\kappa} - \gamma_n(Hn)^{-1/2}\}$. If we have $\gamma_n \leq c_2 H^{1/2} n^{1/2-\kappa}/2$, then the probability bound above gives $\mathrm{P}(\mathcal{M} \subseteq \hat{\mathcal{M}}) \geq 1 - \exp(-c_3 n^{1-2\kappa})$.

Finally, since $q_n = f/p_n$, the requirement on γ_n can be rewritten as $p_n \leq (f/2)\{1 - \Phi(c_2 H^{1/2} n^{1/2-\kappa}/2)\}^{-1}$. Using the fact that $1 - \Phi(x) \leq x^{-1}\exp(-x^2/2)$, this inequality can be satisfied if $p_n \leq f/2\exp(c_2^2 H n^{1-2\kappa}/8)$. Thus the sure screening property holds as long as

$\log(p_n) = O(n^{1-2\kappa})$.

2.8.5 Proof of Theorem 6

We first show that for $j \in \mathcal{M}^c$, $U_j(\beta)$ evaluated at the true β_{0j} can be approximated by a sum of continuous-time martingales, just as it can in a correctly specified Cox regression. We can then appeal to an Edgeworth expansion by Gu (1992) to conclude.

By Theorem 2 we know that $\beta_{0j} = 0$ for $j \in \mathcal{M}^c$. Thus we can rewrite (2.4) as

$$U_j(\beta_{0j}) = \sum_{i=1}^{n} \int_0^\tau \left\{ Z_{ij} - \frac{n^{-1}\sum_l Z_{lj} Y_l(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_l}}{n^{-1}\sum_l Y_l(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_l}} \right\} dN_i(x) + \tag{2.18}$$

$$\sum_{i=1}^{n} \int_0^\tau \left\{ \frac{n^{-1}\sum_l Z_{lj} Y_l(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_l}}{n^{-1}\sum_l Y_l(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_l}} - \frac{n^{-1}\sum_l Z_{lj} Y_l(x)}{n^{-1}\sum_l Y_l(x)} \right\} dN_i(x) \tag{2.19}$$

$$= \sum_{i=1}^{n} \int_0^\tau \left\{ Z_{ij} - \frac{n^{-1}\sum_l Z_{lj} Y_l(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_l}}{n^{-1}\sum_l Y_l(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_l}} \right\} dM_i(x) + \tag{2.20}$$

$$n^{-1} \sum_{l=1}^{n} \int_0^\tau \left\{ \frac{Z_{lj} Y_l(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_l}}{n^{-1}\sum_l Y_l(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_l}} - \frac{Z_{lj} Y_l(x)}{n^{-1}\sum_l Y_l(x)} \right\} \sum_i^n dN_i(x), \tag{2.21}$$

where $M_i(x) = N_i(x) - \int^x Y_i(t)\lambda_0(t) e^{\boldsymbol{\alpha}_0^T \mathbf{Z}_i} dt$ is a continuous martingale in x.

Now let $S_m = \sum_{l=1}^{m} \xi_l$, where

$$\xi_l = \left\{ \frac{Z_{lj} Y_l(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_l}}{n^{-1}\sum_l Y_l(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_l}} - \frac{Z_{lj} Y_l(x)}{n^{-1}\sum_l Y_l(x)} \right\} \sum_i^n dN_i(x). \tag{2.22}$$

Note that $\mathrm{E}(\xi_m \mid S_{m-1}) = \mathrm{E}\{\mathrm{E}(\xi_m \mid S_{m-1}, \mathbf{Z}_m) \mid S_{m-1}\}$, and

$$\mathrm{E}(\xi_m \mid S_{m-1}, \mathbf{Z}_m) = Z_{mj} \mathrm{E}\left[\left\{ \frac{Y_m(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_m}}{n^{-1}\sum_l Y_l(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_l}} - \frac{Y_m(x)}{n^{-1}\sum_l Y_l(x)} \right\} \sum_i^n dN_i(x) \middle| S_{m-1}, \mathbf{Z}_m \right]. \tag{2.23}$$

Given S_{m-1}, the conditional expectation on the right-hand side above is a random variable in $Z_{mk}, k \in \mathcal{M}$ only, and by Assumption 12 is independent of $Z_{mj}, j \in \mathcal{M}^c$. Since $\mathrm{E}(Z_{mj} \mid S_{m-1}) = \mathrm{E}(Z_{lj}) = 0$ by Assumption 10, we find that $\mathrm{E}(\xi_m \mid S_{m-1}) = 0$, implying that S_m is

a discrete martingale in m. Then when $m = n$, by the inequality of Dharmadhikari et al. (1968) we have that $\mathrm{E}(|n^{-1}S_n|^p) = Dn^{-p/2}$ for $p \geq 2$, where we can show that D does not depend on j.

We have shown that for $j \in \mathcal{M}^c$,

$$U_j(\beta_{0j}) = \sum_{i=1}^{n} \int_0^\tau \left\{ Z_{ij} - \frac{n^{-1}\sum_l Z_{lj} Y_l(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_l}}{n^{-1}\sum_l Y_l(x) e^{\boldsymbol{\alpha}_o^T \mathbf{Z}_l}} \right\} dM_i(x) + n^{-1} S_n, \tag{2.24}$$

where $n^{-1}S_n$ satisfies the same conditions as $R_{1,n}$ in (4.3) of Gu (1992). We can therefore extend the proof of Theorem 2.1 in Gu (1992) to show that

$$\sup_x \left| \mathrm{P}\{I_j(\hat{\beta}_j)^{1/2} |\hat{\beta}_j| \geq x\} - \Phi(x) \right| \leq c_4 n^{-1/2}, \tag{2.25}$$

where c_4 does not depend on j. Then (2.6) implies

$$\mathrm{E}\left(\frac{\left|\hat{\mathcal{M}} \cap \mathcal{M}^c\right|}{|\mathcal{M}^c|} \right) \leq \frac{1}{p_n - s_n} \sum_{j \in \mathcal{M}^c} \left[2\{1 - \Phi(\gamma_n)\} + c_4 n^{-1/2} \right]. \tag{2.26}$$

The result follows if we choose $\gamma_n = \Phi^{-1}(1 - q_n/2)$.

Sure screening for estimating equations in ultra-high dimensions

Sihai Dave Zhao and Yi Li

Department of Biostatistics
Harvard School of Public Health

3.1 Introduction

Modern high-throughput experiments are producing high-dimensional datasets with extremely large numbers of covariates. Traditional regression modeling strategies work poorly in such situations, leading to recent interest in regularized regression methods such as the lasso (Tibshirani, 1996), the Dantzig selector (Candès and Tao, 2007), and SCAD (Fan and Li, 2001). These procedures can perform well in estimation and prediction even when the number of covariates p_n is larger than the sample size n, where here we are allowing p_n to grow with n. However, when p_n is extremely large compared to n, these methods can become inaccurate and computationally infeasible (Fan and Lv, 2008). Thus there is a need for methods to quickly screen out unimportant covariates before using regularization methods.

A number of screening strategies have so far been proposed, and choosing which one to use depends on what model we believe is most suitable for the data. Under the ordinary linear model, Fan and Lv (2008) proposed a procedure with the sure screening property, where the covariates retained after screening will contain the truly important covariates with probability approaching one, even in the ultra-high-dimensional realm where p_n grows exponentially with n. Fan and Song (2010) and Zhao and Li (2012) subsequently proposed procedures that maintain this property for generalized linear models and the Cox model, respectively. Screening methods have also been proposed for nonparametric additive models (Fan et al., 2011), linear transformation models (Li et al., 2011), and single-index hazard models (Gorst-Rasmussen and Scheike, 2011).

In a recent development, Zhu et al. (2011) proposed a screening method valid for any single-index model, a class so large that their screening procedure is nearly model-free. They used a new measure of dependence which can detect a wide variety of functional relationships between the covariates and the outcome, and proved that their method has the sure screening property for any single-index model. They also showed in simulations that it could significantly outperform model-based screening methods when the models were

incorrectly specified.

On the other hand, model-based screening can have greater power to detect important covariates, a consequence of the bias-variance tradeoff. However, there are often situations where we wish to use some model other than the ones mentioned above. For example, studies involving clustered observations, missing data, or censored outcomes are frequently encountered in genomic medicine, and are often analyzed with more complicated regression models for which no screening methods have yet been developed. In theory it is not difficult to propose a screening procedure for any given model: fit p_n marginal regressions, one for each covariate, and retain those covariates with the largest marginal estimates, in absolute value. But fitting p_n marginal regressions can still be time-consuming, especially if p_n is very large and the fitting procedure is slow, and theoretical properties such as sure screening must still be studied on a case-by-case basis.

In this paper we propose EEScreen, a unified approach to screening which can be used with any statistical model that can be fit using estimating equations. This is convenient because estimating equations are frequently used to analyze the previously mentioned correlated, missing, or censored data situations. EEScreen is also fundamentally different from most other screening procedures in that it only requires evaluating p_n estimating equations at a fixed parameter value, rather than solving for p_n marginal regressione estimates, making it exceedingly computationally convenient. We prove theoretical results about the screening properties of EEScreen that hold for any model that can be fit using U-statistic-based estimating equations.

Furthermore, because we can design estimating equations to incorporate more or fewer modeling assumptions, we can use our EEScreen framework to span the range between model-based and model-free screening. In particular, we show that EEScreen can provide a screening method very similar to that of Zhu et al. (2011) when used with a particular estimating equation. This estimating equation actually cannot be used for estimation in practice because it involves unknown parameters, but interestingly can still be used to derive

a useful screening procedure.

Finally, when covariates are highly correlated, Fan and Lv (2008) suggested an iterative version of their screening procedure, which they found to outperform marginal screening in some cases. In this paper we provide an iterative version of EEScreen (iEEScreen), and we also demonstrate a novel connection between iEEScreen and EEBoost, a recently proposed boosting algorithm for estimation and variable selection in estimating equations (Wolfson, 2011). This connection may provide a means for a theoretical analysis of iterative screening methods, something which so far has been difficult to study.

We introduce EEScreen in Section 3.2, where we also give some examples, establish its theoretical properties, and briefly discuss how to choose the number of covariates to retain after screening. We derive a new screening method similar to that of Zhu et al. (2011) in Section 3.3, and discuss iEEScreen in Section 3.4. We conduct a thorough simulation study in Section 3.5, using two different estimating equations, before applying our methods to analyze an issue in multiple myeloma in Section 3.6. We conclude with a discussion in Section 3.7, and provide proofs in the Appendix.

3.2 EEScreen: sure screening for estimating equations

3.2.1 Method

Let $Y_i = (Y_{i1}, \ldots, Y_{iK_i})^T$ be a $K_i \times 1$ outcome vector and $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iK_i})^T$ be a $K_i \times p_n$ matrix of covariates for units $i = 1, \ldots, n$. Then let $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^T$ be a $\sum_i K_i \times 1$ vector and $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T$ be a $\sum_i K_i \times p_n$ matrix. Assuming some regression model, we can construct a $p_n \times 1$ estimating equation $\mathbf{U}(\boldsymbol{\beta})$ that depends on $\mathbf{Y}_i$ and $\mathbf{X}_i$ such that $\mathrm{E}\{\mathbf{U}(\boldsymbol{\beta}_0)\} = \mathbf{0}$, where $\boldsymbol{\beta}_0$ is the true $p_n \times 1$ parameter vector. Let the set of true regression parameters $\mathcal{M} = \{j : \beta_{0j} \neq 0\}$ have size $|\mathcal{M}| = s_n$, where β_{0j} is the j^{th} component of $\boldsymbol{\beta}_0$.

It is commonly assumed that s_n is small and fixed or growing slowly. When $p_n < n$, $\boldsymbol{\beta}_0$ is estimated by finding the $\hat{\boldsymbol{\beta}}$ such that $\mathbf{U}(\hat{\boldsymbol{\beta}}) = 0$, but when $p_n > n$ there are an infinite number of solutions for $\hat{\boldsymbol{\beta}}$, in which case regularized regression is used (Fu, 2003; Johnson et al., 2008; Wolfson, 2011). However, when p_n is much greater than n, these methods can lose accuracy and be too computationally demanding, hence the need for screening methods to quickly reduce p_n.

Most previously proposed screening methods proceed by fitting p_n regression models, one covariate at a time, to get p_n marginal estimates $\hat{\alpha}_j$. They then retain the covariates with $|\hat{\alpha}_j|$ above some threshold. This is akin to conducting p_n Wald tests, though without standardizing the $\hat{\alpha}_j$ by their variances. However, in the case of estimating equations, even this procedure can be time-consuming if p_n is large or $\mathbf{U}$ is cumbersome to fit.

Here, instead of marginal Wald tests, we construct marginal score tests for the β_{0j} using $\mathbf{U}$. To motivate our procedure, we first consider the case where the marginal model is correct for β_{01}. In other words, $\beta_{01} \neq 0$ while $\beta_{0j} = 0$ for all $j \neq 1$. Then $\mathrm{E}[\mathbf{U}\{(\beta_{01}, 0, \ldots, 0)\}] = \mathbf{0}$, so that each component of $\mathbf{U}$ is a valid estimating equation for β_{01}. This implies that each component of $\mathbf{U}(\mathbf{0})$ is the numerator of a score test for the null hypothesis $\beta_{01} = 0$. If the marginal model is correct for β_{01}, then to achieve sure screening we must reject the score test. Therefore we use as our screening statistic the component of $\mathbf{U}(\mathbf{0})$ that gives the most powerful test, which we denote $U_1(\mathbf{0})$. For each j, we can identify the component $U_j(\mathbf{0})$ of $\mathbf{U}(\mathbf{0})$ that is most powerful for testing $\beta_{0j} = 0$ under the marginal model that β_{0j} is the only non-zero parameter. In many situations the first component of $\mathbf{U}(\mathbf{0})$ will be associated with β_{01}, the second with β_{02}, and so on. When this is not the case, we can follow the construction above to relabel the components of $\mathbf{U}(\mathbf{0})$ appropriately.

We propose using the relabeled $U_j(\mathbf{0})$ as surrogate measures of association between the outcome and the j^{th} covariate, after first standardizing the covariates to have equal variances. Instead of just taking the numerators of the score tests we could divide each $U_j(\mathbf{0})$ by an estimate of its standard deviation, but this would add computational complexity to our

procedure, and even without doing so we will be able to achieve good results and prove finite-sample performance guarantees. One advantage to using score tests is that they do not require parameter estimation and so are more computationally convenient than performing p_n marginal regressions. Furthermore, this framework will also allow us to give a unified treatment of the theoretical results for a large class of estimating equations.

Specifically, we propose the following screening procedure:

1. Standardize the p_n covariates to have variance 1.
2. For the j^{th} parameter identify the marginal estimating equations U_j as described above.
3. Set a threshold γ_n.
4. Retain the parameters $\{j : |U_j(\mathbf{0})| \geq \gamma_n\}$.

We will denote the set of retained parameters by $\hat{\mathcal{M}}$. Note that this procedure only requires evaluating p_n estimating equations at $\mathbf{0}$, which can be computed very quickly. The convenience of score tests, however, comes at the price of ambiguity in the proper treatment of nuisance parameters, such as the intercept term in a regression model. Without loss of generality, let β_{01} be the intercept term. We can first fit the intercept without any covariates in the model to get an estimate $\hat{\beta}_{01}$. This only needs to be done once, since $\hat{\beta}_{01}$ will remain the same for each U_j. We then screen by evaluating each U_j at $\boldsymbol{\eta} = (\hat{\beta}_{01}, \mathbf{0})$ instead of at $\mathbf{0}$.

Our score test idea was motivated by the EEBoost algorithm (Wolfson, 2011), a boosting procedure for estimating equations which uses components of the estimating equation $\mathbf{U}$ as a surrogate measure of association. We therefore refer to our method as EEScreen, and we will draw more connections between EEScreen and EEBoost in Section 3.4.

3.2.2 Examples

Here we provide some examples of EEScreen for various estimating equations, assuming throughout that $\mathrm{E}(\mathbf{X}_i) = \mathbf{0}$ and $\mathrm{var}(X_{ij}) = 1$. For the linear model with $K_i = 1$, the usual linear regression score equation is $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$, so $\mathbf{U}(\mathbf{0}) = \mathbf{X}^T\mathbf{Y}$. Under the marginal model that $\beta_{0j'}$ is the only non-zero parameter, the j^{th} component of $\mathrm{E}\{\mathbf{U}(\mathbf{0})\}$ equals $\mathrm{cor}(X_{ij}, X_{ij'})\beta_{0j'}$, where X_{ij} is the j^{th} component of the i^{th} covariate vector. Clearly this is maximized when $j = j'$ for any value of $\beta_{0j'}$, so the component of $\mathbf{U}(\mathbf{0})$ that gives the most powerful test is $U_{j'}(\mathbf{0})$. EEScreen then retains the parameters $\{j : |\sum_i X_{ij}Y_i| \geq \gamma_n\}$. Note that this is equivalent to the original screening procedure proposed by Fan and Lv (2008).

Under the Cox model, when $K_i = 1$ with survival outcomes, let T_i be the survival time, C_i the censoring time, $Y_i = \min(T_i, C_i)$, and $\delta_i = I(T_i \leq C_i)$. The Cox model score equation is

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int \left\{ \mathbf{X}_i - \frac{\sum_{i=1}^{n} \mathbf{X}_i \tilde{Y}_i(x) \exp(\mathbf{X}_i^T\boldsymbol{\beta})}{\sum_{i=1}^{n} \tilde{Y}_i(x) \exp(\mathbf{X}_i^T\boldsymbol{\beta})} \right\} d\tilde{N}_i(x), \tag{3.1}$$

where $\tilde{N}_i(x) = I(T_i \leq x, \delta_i)$ is the observed failure process and $\tilde{Y}_i(x) = I(Y_i \geq x)$ is the at-risk process. Under the marginal model that $\beta_{0j'}$ is the only non-zero parameter, Gorst-Rasmussen and Scheike (2011) show that the largest component of the limiting estimating equation evaluated at $\mathbf{0}$ is found for the j that maximizes $\int \mathrm{cor}\{X_{ij}, F(t \mid X_{ij'})\}$, where $F(t \mid X_{ij'})$ is the distribution function of T_i, conditional on $X_{ij'}$. Thus the component of $\mathbf{U}(\mathbf{0})$ that gives the most powerful test is again the j'^{th} component. EEScreen then retains the parameters

$$\left[j : \left| \sum_{i=1}^{n} \int \left\{ X_{ij} - \frac{\sum_{i=1}^{n} X_{ij} \tilde{Y}_i(x)}{\sum_{i=1}^{n} \tilde{Y}_i(x)} \right\} d\tilde{N}_i(x) \right| \geq \gamma_n \right]. \tag{3.2}$$

This is exactly the screening statistic of Gorst-Rasmussen and Scheike (2011). This example illustrates the computational advantages that EEScreen can enjoy. Zhao and Li (2012) proposed screening for the Cox model based on fitting marginal Cox regressions, which requires p_n applications of the Newton-Raphson algorithm. In contrast, Gorst-Rasmussen

and Scheike (2011) and EEScreen only require evaluating the $U_j(\mathbf{0})$.

The ordinary linear model and the Cox model have already been studied in the screening literature, but EEScreen is most useful for models for which no screening procedures exist yet. In Sections 3.5 we study its performance on two such models: a t-year survival model (Jung, 1996) and the accelerated failure time model (Tsiatis, 1996; Jin et al., 2003).

3.2.3 Theoretical properties

One advantage of our EEScreen framework is that we can provide very general theoretical guarantees on its screening performance that hold for a large class of models, without needing to study each model on a case-by-case basis. We require three assumptions on the marginal estimating equations U_j to prove that EEScreen has the sure screening property, where the probability that the retained parameters $\hat{\mathcal{M}}$ contains the true parameters $\mathcal{M}$ approaches 1. Let the expected full estimating equations be denoted $\mathbf{u}(\boldsymbol{\beta}) = \mathrm{E}\{\mathbf{U}(\boldsymbol{\beta})\}$, so that the expected marginal estimating equations are $u_j(\boldsymbol{\beta})$.

Assumption 13 *Let $\mathbf{X}_{ij}$ be the $K_i \times 1$ vector of the j^{th} covariate for the i^{th} unit. Each estimating equation U_j has the form*

$$U_j(\boldsymbol{\beta}) = \binom{n}{m}^{-1} \sum_{1 \le i_1 < \ldots < i_m} h_j\{\boldsymbol{\beta}; (\mathbf{Y}_{i_1}, \mathbf{X}_{i_1}), \ldots, (\mathbf{Y}_{i_m}, \mathbf{X}_{i_m})\} \tag{3.3}$$

for all j, where $n \ge m$ and h_j is a real-valued kernel function that depends on β and is symmetric in the $(\mathbf{Y}_{i_1}, \mathbf{X}_{i_1}), \ldots, (\mathbf{Y}_{i_m}, \mathbf{X}_{i_m})$.

Assumption 14 *There exist some constants $b > 0$ and $\Sigma^2 > 0$ such that for all j, $|U_j(\mathbf{0}) - u_j(\mathbf{0})| \le b$ and $\mathrm{var}[h_j\{\mathbf{0}; (\mathbf{Y}_{i_1}, \mathbf{X}_{i_1}), \ldots, (\mathbf{Y}_{i_m}, \mathbf{X}_{i_m})\}] \le \Sigma^2$.*

Assumption 13 requires that each U_j be a U-statistic of order m, which encompasses a large number of important estimating equations. Assumption 14 amounts to conditions on the moments of the U_j, and they can often be satisfied by assuming bounded outcomes and covariates. These conditions are necessary for stating a Bernstein-type inequality for the U_j, which gives the probability bounds in Theorems 7 and 8. They can therefore be relaxed as long as there exists a similar probability inequality for U_j. For example, Bernstein-type inequalities exist for martingales (van de Geer, 1995), which would allow U_j to be the Cox model score equations.

Assumption 15 *There exists some constant* $c_1 > 0$ *such that* $\min_{j \in \mathcal{M}} |u_j(\mathbf{0})| \geq c_1 [n/m]^{-\kappa}$ *with* $0 < \kappa < 1/2$*, where* m *is defined in Assumption 13 and* $[n/m]$ *is the largest integer less than* n/m*.*

Assumption 15 is an assumption on the marginal signal strengths of the covariates in $\mathcal{M}$. In EEScreen these signals are quantified by the $u_j(\mathbf{0})$, and Assumption 15 requires them to be of at least a certain order so that they are detectable given a sample size n. An assumption of this type is always needed in a theoretical analysis of a screening procedure. For example, in the generalized linear model setting, our Assumption 15 is exactly equivalent to the assumption of Fan and Song (2010) that the magnitude of the covariance between $\mathrm{E}(\mathbf{Y}_i \mid \mathbf{X}_i)$ and the j^{th} covariate be of order $n^{-\kappa}$. Since EEScreen is similar to conducting p_n score tests, Assumption 15 is similar to requiring that the expected value of the marginal score test statistic for $j \in \mathcal{M}$ be of a certain order. As previously mentioned, we could standardize the screening statistic $|u_j(\mathbf{0})|$ by its variance, in which case the score test analogy would be exact. It is very reasonable to use the marginal score test statistic as a proxy for the marginal association of the covariates.

Under these assumptions, we can show that EEScreen possesses the sure screening property.

Theorem 7 *Under Assumptions 13–15, if* $\gamma_n = c_1[n/m]^{-\kappa}/2$ *for* $0 < \kappa < 1/2$, *with* m *defined in Assumption 13, then*

$$\mathrm{P}(\mathcal{M} \subseteq \hat{\mathcal{M}}) \geq 1 - 2s_n \exp\left\{-\frac{c_1^2[n/m]^{1-2\kappa}/4}{2\Sigma^2 + bc_1[n/m]^{-\kappa}/3}\right\}, \tag{3.4}$$

with Σ^2 *and* b *defined in Assumption 14.*

Theorem 7 guarantees that all important covariates will be retained by EEScreen with high probability. Similar to previous work, we find that this probability bound depends only on s_n and not on p_n. The bound also depends on m, the order of the U-statistic, so that EEScreen may not perform as well for larger m. Theorem 7 is almost an immediate consequence of properties of U-statistics, and the simplicity of the proof is due to the fact that EEScreen uses score tests instead of Wald tests. We therefore do not need to estimate any parameters, nor prove probability inequalities for those estimates, which is a major source of technical difficulty in previous work on screening.

Theorem 7 is most useful if the size of the $\hat{\mathcal{M}}$ produced by EEScreen is small. In other words, we hope that $\hat{\mathcal{M}}$ does not contain too many false positives. With two more assumptions, we can provide a bound on $|\hat{\mathcal{M}}|$ that holds with high probability.

Assumption 16 *The expected full estimating equation* $\mathbf{u}(\boldsymbol{\beta})$ *is differentiable with respect to* $\boldsymbol{\beta}$. *Let the negative Jacobian* $-\partial\mathbf{u}/\partial\boldsymbol{\beta}$ *be denoted* $\mathbf{i}(\boldsymbol{\beta})$.

Assumption 17 *There exists some constant* $c_2 > 0$ *such that* $\|\boldsymbol{\beta}_0\|_2 \leq c_2$.

Assumption 16 can hold even if the sample estimating equation $\mathbf{U}$ is nondifferentiable. Assumption 17 merely requires that there exist an upper bound on the size of the true $\boldsymbol{\beta}_0$ that does not grow with n, which is a reasonable condition.

Theorem 8 *Under Assumptions 13–17, if* $\gamma_n = c_1[n/m]^{-\kappa}/2$ *as in Theorem 7, then*

$$\mathrm{P}\left[|\hat{\mathcal{M}}| \leq \frac{16c_2^2\sigma_{\max}^{*2}}{c_1^2[n/m]^{-2\kappa}}\right] \geq 1 - 2p_n \exp\left\{-\frac{c_1^2[n/m]^{1-2\kappa}/16}{2\Sigma^2 + bc_1[n/m]^{-\kappa}/6}\right\}, \tag{3.5}$$

where Σ^2 *and* b *are defined in Assumption 14 and* $\sigma_{\max}^* = \sup_{0<t<1} \sigma_{\max}\{\mathbf{i}(t\boldsymbol{\beta}_0)\}$*, where* $\sigma_{\max}(\mathbf{A})$ *denotes the largest singular value of the matrix* $\mathbf{A}$*.*

Like Theorem 7, Theorem 8 is also almost a simple consequence of properties of U-statistics. Theorem 8 provides a finite-sample probability bound on $|\hat{\mathcal{M}}|$, but asymptotically we would need assumptions on $\mathbf{i}(\boldsymbol{\beta}^*)$ to guarantee that $\sigma_{\max}^*$ will not increase too quickly. In particular, if $\sigma_{\max}^*$ increased only polynomially in n, $|\hat{\mathcal{M}}|$ would increase polynomially. At the same time, the probability that the bound holds tends to one even if $\log p_n = o([n/m]^{1-2\kappa})$, so the false positive rate would decrease quickly to zero with probability approaching one even in ultra-high dimensions. A similar phenomenon was found by Fan et al. (2011).

The presence of $\sigma_{\max}^*$ in Theorem 8 reflects the dependence of $|\hat{\mathcal{M}}|$ on the degree of collinearity of our data. For general estimating equations, collinearity not only depends on the design matrix, but also varies across the parameter space. For example, Mackinnon and Puterman (1989) and Lesaffre and Marx (1993) showed that generalized linear models can be collinear even if their design matrices are not, and vice versa. In our situation, we are concerned with collinearity along the line segment between $\boldsymbol{\beta}_0$ and $\mathbf{0}$. Note that because $\sigma_{\max}^*$ depends only on $\mathbf{i}$, $\boldsymbol{\beta}_0$, and $\mathbf{0}$, which are all nonrandom quantities, $\sigma_{\max}^*$ is nonrandom as well.

3.2.4 Choosing γ_n

Theorems 7 and 8 specify optimal rates for γ_n, and a number of methods have been proposed for choosing γ_n in practice. Fan and Lv (2008) suggested choosing γ_n such that $|\hat{\mathcal{M}}| = n - 1$ or $n/\log n$. Because these values are hard to interpret, Zhao and Li (2012) showed that

γ_n is related to the expected false positive rate of screening. Zhu et al. (2011) also recently proposed a thresholding method based on adding artificial auxiliary variables, and provided a bound relating the number of added variables to the probability of including an unimportant covariate. These methods offer more interpretable ways of choosing how many covariates to retain with EEScreen. A related strategy is to set a desired false discovery rate. Bunea et al. (2006) showed that FDR methods can achieve the sure screening property in the ordinary linear model, and Sarkar (2004) proposed an FDR method than can also control the false negative rate. It would be interesting to pursue this type of idea for EEScreen.

In practice, however, we are often concerned with the prediction error of the estimator obtained by fitting a regularized regression method after EEScreen. If we used the methods above we would still need to choose a false positive rate or false discovery rate, but so far it is not clear what choices would give optimal prediction. In this case another option is to retain different numbers of covariates, fit the regularized regression for each screened model $\hat{\mathcal{M}}$, and select the $\hat{\mathcal{M}}$ that gives the lowest cross-validated estimate of prediction error. This is the approach we take in Section 3.6, where we use EEScreen to analyze data from a multiple myeloma clinical trail.

3.3 Model-free screening

Zhu et al. (2011) recently proposed a screening statistic that can achieve sure screening for any single-index model. Specifically, for a completely observed response $\tilde{Y}_i$ and a p-dimensional covariate vector $\mathbf{X}_i$, they assumed that $F(y \mid \mathbf{X}_i) = F_0(y \mid \mathbf{X}_i^T\boldsymbol{\beta}_0)$, where $F(y \mid \mathbf{X}_i) = \mathrm{P}(\tilde{Y}_i < y \mid \mathbf{X}_i)$ and F_0 is some distribution function that depends on $\mathbf{X}_i$ only through the index $\mathbf{X}_i^T\boldsymbol{\beta}_0$, so that $j \in \mathcal{M}$ if and only if $\beta_{0j} \neq 0$. This is a very mild assumption that holds for a large class of models, making the screening method of Zhu et al. (2011) almost model-free.

To simplify things, they assumed that $\mathrm{E}(\mathbf{X}_i) = \mathbf{0}$ and $\mathrm{var}(\mathbf{X}_i) = \mathbf{I}_{p_n}$, where $\mathbf{I}_{p_n}$ is the $p_n \times p_n$ identity matrix. They quantified the marginal relationship between the covariates and an outcome y by using the novel statistic

$$\boldsymbol{\Omega}(y) = \mathrm{E}\{\mathbf{X}_i F(y \mid \mathbf{X}_i)\} = \mathrm{cov}\{\mathbf{X}_i, F(y \mid \mathbf{X}_i)\} = \mathrm{cov}\{\mathbf{X}_i, I(\tilde{Y}_i < y)\}. \tag{3.6}$$

Intuitively, the covariance between X_{ij} and $F(\tilde{Y}_i \mid \mathbf{X}_i)$, where X_{ij} is the j^{th} component of $\mathbf{X}_i$, should be large in magnitude if $j \in \mathcal{M}$. They therefore used $\omega_j = \mathrm{E}\{\Omega_j(\tilde{Y}_i)^2\}$ as a measure of marginal association, where $\Omega_j(y)$ is the j^{th} component of $\boldsymbol{\Omega}(y)$, leading to the screening statistic

$$\tilde{\omega}_j = n^{-1} \sum_{k=1}^{n} \left\{ n^{-1} \sum_{i=1}^{n} X_{ij} I(\tilde{Y}_i < \tilde{Y}_k) \right\}^2. \tag{3.7}$$

This derivation of the screening procedure of Zhu et al. (2011) makes no mention of estimation of $\boldsymbol{\beta}_0$, making it seemingly irreconcilable with our EEScreen, which requires an estimating equation. However, we can actually show that EEScreen, combined with a particular estimating equation, leads to a very similar screening procedure. This further illustrates the flexibility and wide applicability of our proposed screening strategy.

Note that conditional on $\mathbf{X}_i$ and $\mathbf{X}_k$, $F_0(\tilde{Y}_i \mid \mathbf{X}_i^T \boldsymbol{\beta}_0)$ and $F_0(\tilde{Y}_k \mid \mathbf{X}_k^T \boldsymbol{\beta}_0)$ are independent and identically distributed uniform random variables. Therefore, we know that

$$\mathrm{P}\left\{ F_0(\tilde{Y}_i \mid \mathbf{X}_i^T \boldsymbol{\beta}_0) < F_0(\tilde{Y}_k \mid \mathbf{X}_k^T \boldsymbol{\beta}_0) \right\} = \tag{3.8}$$

$$\mathrm{E}\left[\mathrm{P}\left\{ F_0(\tilde{Y}_i \mid \mathbf{X}_i^T \boldsymbol{\beta}_0) < F_0(\tilde{Y}_k \mid \mathbf{X}_k^T \boldsymbol{\beta}_0) \mid \mathbf{X}_i, \mathbf{X}_k \right\} \right] = \frac{1}{2}. \tag{3.9}$$

This fact can be used to construct the marginal estimating equations. Consider

$$\mathbf{U}(\boldsymbol{\beta}) = n^{-2} \sum_{k=1}^{n} \sum_{i=1}^{n} \mathbf{X}_i \left[I\{F_0(\tilde{Y}_i \mid \mathbf{X}_i^T \boldsymbol{\beta}) < F_0(\tilde{Y}_k \mid \mathbf{X}_k^T \boldsymbol{\beta})\} - \frac{1}{2} \right]. \tag{3.10}$$

Since $\mathrm{E}\{\mathbf{U}(\boldsymbol{\beta}_0)\} = \mathbf{0}$, (3.10) is an unbiased estimating equation for $\boldsymbol{\beta}_0$. Furthermore, it is a U-statistic of order $m = 2$, which is covered by the framework of Section 3.2.3.

It is important to note that (3.10) cannot be implemented in practice, because the functional form of $F_0(y \mid \mathbf{X}^T \boldsymbol{\beta})$ is unknown, yet it is still useful for constructing a screening

procedure. Recall that EEScreen uses the statistic $\mathbf{U}(\mathbf{0})$, and for (3.10),

$$\mathbf{U}(\mathbf{0}) = n^{-2} \sum_{k=1}^{n} \sum_{i=1}^{n} \mathbf{X}_i \left[I\{F_0(\tilde{Y}_i \mid \mathbf{X}_i^T \mathbf{0}) < F_0(\tilde{Y}_k \mid \mathbf{X}_k^T \mathbf{0})\} - \frac{1}{2} \right] \tag{3.11}$$

$$= n^{-2} \sum_{k=1}^{n} \sum_{i=1}^{n} \mathbf{X}_i \left\{ I(\tilde{Y}_i < \tilde{Y}_k) - \frac{1}{2} \right\}, \tag{3.12}$$

because $F_0(y \mid \mathbf{X}_i^T \mathbf{0}) = F_0(y \mid \mathbf{X}_k^T \mathbf{0}) = F_0(y \mid \mathbf{0})$, which is a monotonic function since F_0 is a distribution function. Under the marginal model that $\beta_{0j'}$ is the only non-zero parameter, the j^{th} component of $\mathrm{E}\{\mathbf{U}(\mathbf{0})\}$ is $\mathrm{cor}\{X_{ij}, F(\tilde{Y}_i \mid X_{ij'})\}$. Thus the j'^{th} component of $\mathbf{U}(\mathbf{0})$ gives the most powerful score test, so EEScreen with (3.10) retains parameters

$$\left[j : \left| n^{-2} \sum_{k=1}^{n} \sum_{i=1}^{n} X_{ij} \left\{ I(\tilde{Y}_i < \tilde{Y}_k) - \frac{1}{2} \right\} \right| \geq \gamma_n \right], \tag{3.13}$$

or equivalently,

$$\left\{ j : \left| n^{-2} \sum_{k=1}^{n} \sum_{i=1}^{n} X_{ij} I(\tilde{Y}_i < \tilde{Y}_k) \right| \geq \gamma_n \right\}, \tag{3.14}$$

because the $\mathbf{X}_i$ are standardized to have mean $\mathbf{0}$. In the notation of Zhu et al. (2011), this is equivalent to using $|E\{\Omega_j(\tilde{Y}_i)\}|$ as the screening statistic for the j^{th} covariate, rather than $\mathrm{E}\{\Omega_j(\tilde{Y}_i)^2\}$.

The $\tilde{Y}_i$ may not be fully observed in the presence of censoring. If C_i are the censoring times, let $Y_i = \min(\tilde{Y}_i, C_i)$ and $\delta_i = I(\tilde{Y}_i \leq C_i)$. Then if we assume that the C_i are independent of the $\tilde{Y}_i$ and $\mathbf{X}_i$, we can see that

$$\mathrm{E}\left\{ \frac{\delta_i I(Y_i < Y_k)}{S_C^2(Y_i)} \middle| \mathbf{X}_i, \mathbf{X}_k \right\} = \mathrm{E}\left[\mathrm{E}\left\{ \frac{I(\tilde{Y}_i \leq C_i) I(\tilde{Y}_i \leq C_k) I(\tilde{Y}_i \leq \tilde{Y}_k)}{S_C^2(\tilde{Y}_i)} \middle| \tilde{Y}_i, \mathbf{X}_i, \mathbf{X}_k \right\} \right] \tag{3.15}$$

$$= \mathrm{E}\left[\mathrm{E}\left\{ \frac{S_C^2(\tilde{Y}_k) I(\tilde{Y}_i \leq \tilde{Y}_k)}{S_C^2(\tilde{Y}_i)} \middle| \tilde{Y}_i, \mathbf{X}_i, \mathbf{X}_k \right\} \right] \tag{3.16}$$

$$= \mathrm{E}\{I(\tilde{Y}_k < \tilde{Y}_k) \mid \mathbf{X}_i, \mathbf{X}_k\}, \tag{3.17}$$

where S_C is the survival function of the C_i. If the support of the C_i is less than that of the $\tilde{Y}_i$, the $S_C(Y_i)$ term above could equal 0 for some Y_i. Thus this method of accommodating censoring could cause difficulty if it were used in the estimating equation (3.10) and could

lead to inconsistent estimation of $\boldsymbol{\beta}_0$ (Fine et al., 1998). For simplicity, we will assume here that the support of C_i is greater than or equal to that of $\tilde{Y}_i$.

This then suggests that in the presence of censoring, the screening statistic of Zhu et al. (2011) should become

$$n^{-1}\sum_{k=1}^{n}\left\{n^{-1}\sum_{i=1}^{n}X_{ij}\frac{\delta_i I(Y_i < Y_k)}{\hat{S}_C^2(Y_i)}\right\}^2, \tag{3.18}$$

and the screening statistic derived using EEScreen should become

$$\left|n^{-2}\sum_{k=1}^{n}\sum_{i=1}^{n}X_{ij}\frac{\delta_i I(Y_i < Y_k)}{\hat{S}_C^2(Y_i)}\right|, \tag{3.19}$$

where $\hat{S}_C$ is the Kaplan-Meier estimate of S_C. This illustrates that the EEScreen framework is flexible enough to allow us to derive something similar to the approach of Zhu et al. (2011), which was originally motivated by very different considerations. It also suggests that EEScreen can provide a sensible screening procedure for a particular model, such as the single-index model, even if the associated estimating equation (3.10) is not implementable in practice.

3.4 iEEScreen

Though the simplicity of EEScreen and related screening procedures is appealing, if the covariates are highly correlated, then in finite samples these univariate screening methods may not be able to achieve sure screening without incurring a large number of false positives. To address this issue, Fan and Lv (2008) and Fan et al. (2009) proposed iterative screening, where the general idea is as follows. Below, $\mathcal{M}_l$ and $\mathcal{A}_l$ denote sets of covariate indices. In other words, $\mathcal{M}_l, \mathcal{A}_l \subseteq \{1, \ldots, p_n\}$.

1. Set $\mathcal{M}_0$ to be the empty set.

2. For $l = 1 : L$,

(a) controlling for the variables in $\mathcal{M}_{l-1}$, screen the remaining covariates

(b) select a set $\mathcal{A}_l$ of the most important of these covariates

(c) use a multivariate variable selection method, such as lasso or SCAD, on the covariates in $\mathcal{M}_{l-1} \cup \mathcal{A}_l$ to get a reduced set $\mathcal{M}_l$

We can adapt these ideas to develop an iterative version of EEScreen, which we will call iEEScreen. However, to operationalize iEEScreen and iterative screening algorithms in general, we must first specify a number of parameters, such as how large $|\mathcal{A}_l|$ and $|\mathcal{M}_l|$ should be, what multivariate variable selection procedure to use, and how many iterations to run. Fan et al. (2011) recommended choosing the $\mathcal{A}_l$ using a permutation-based procedure, and the $\mathcal{M}_l$ using a SCAD-type variable selector (Fan and Li, 2001) with cross-validation. Their iterations stop when either $|\mathcal{M}_l| > |\mathcal{A}_1|$, or $\mathcal{M}_l = \mathcal{M}_{l-1}$. These are sensible choices, but the many different layers of this procedure make it difficult to analyze.

Instead, here we will show that the EEBoost method of Wolfson (2011), viewed as a variable selector rather than an estimation procedure, can actually be thought of as a version of iEEScreen. By linking iterative screening and boosting, we embed iEEScreen in the theoretical framework already developed for EEBoost and other boosting methods. In the future, this theoretical framework could in turn be applied to analyze the properties of iterative screening.

We first briefly describe the EEBoost algorithm (Wolfson, 2011). For some small $\epsilon > 0$ and the full estimating equation $\mathbf{U}$,

1. Set $\boldsymbol{\beta}^{(0)} = \mathbf{0}$.

2. For $t = 1 : T$,

   (a) compute $\boldsymbol{\Delta} = |\mathbf{U}(\boldsymbol{\beta}^{(t-1)})|$

   (b) identify $j_t = \text{argmax}_j \, \Delta_j$, where Δ_j is the j^{th} component of $\boldsymbol{\Delta}$

(c) set $\beta_{j_t}^{(t)} = \beta_{j_t}^{(t-1)} - \epsilon \cdot \text{sign}(\Delta_{j_t})$, where $\beta_{j_t}^{(t)}$ is the j_t^{th} component of $\boldsymbol{\beta}^{(t)}$

Here, T serves as the regularization parameter, and for a given T only a certain number of $\boldsymbol{\beta}_{j_t}^{(t)}$ will have been updated from their initial values of zero, effecting variable selection. Wolfson (2011) recommends choosing ϵ in the range [0.001,0.05], and T can be chosen with some tuning procedure.

To express EEBoost as an iterative version of EEScreen, note that at the beginning of EEBoost, Δ_j corresponds to the screening statistic $|U_j(0)|$ used in EEScreen. Evaluating $\mathbf{U}$ at subsequent $\boldsymbol{\beta}^{(t-1)}$ is a way of controlling for the variables that have already been selected into the model by EEBoost, which is step 2(a) of iterative screening. In particular, for $i = 0, 1, \ldots$ define t_i such that $\|\boldsymbol{\beta}^{(t_i)}\|_0 \neq \|\boldsymbol{\beta}^{(t_i+1)}\|_0$. In other words, t_0 is the first time that the number of nonzero components of $\boldsymbol{\beta}^{(t)}$ changes, t_1 is the second time this happens, and so on. Then looking back at the iterative screening algorithm, for $l = 1, \ldots, L$ we can identify $\mathcal{M}_{l-1}$ to be $\{j : \beta_j^{(t_{l-1})} \neq 0\}$, $\mathcal{A}_l$ to be $\{j_{t_l}\}$, and $\mathcal{M}_l$ as being obtained by running EEBoost for t_l iterations starting from the covariates in $\mathcal{M}_{t_l-1} \cup \mathcal{A}_l$. We can choose L by tuning EEBoost with a generalized cross-validation-type criterion. We will thus implement iEEScreen using the EEBoost algorithm.

In the remainder of this paper we study the effects of using EEScreen and iEEScreen as preprocessing steps before fitting regularized regression models. In particular, we will use EEBoost to fit the regressions, for two reasons. First, we would like to compare the effects of retaining different numbers of covariates after screening, from keeping only one or two covariates to keeping tens of thousands. Therefore we require a regularization method for estimating equations that can handle an arbitrarily large number of covariates. Second, in Section 3.5.2 we study a discrete estimating equation, so we require a regularization method which works well in that situation. To our knowledge, EEBoost is the only procedure that meets both of these criteria.

However, this leads to a unique problem. We would naturally like to compare the effects of using EEScreen versus iEEScreen. But a careful inspection of the EEBoost algorithm reveals that running EEBoost twice, in other words first selecting covariates using EEBoost, and then using only those covariates in another instance of EEBoost, is actually identical to using EEBoost only once. This means that screening with the version of iEEScreen described in this section has no effect if EEBoost is then used for model-fitting. This behavior is different from, say, the lasso, where running two iterations of the lasso has been termed the relaxed lasso (Meinshausen, 2007) and can give different results from the regular lasso. Therefore while we will be able to compare the variable selection properties of EEScreen and iEEScreen in simulations, where we will know the true model, we will not be able to compare EEScreen+EEBoost versus iEEScreen+EEBoost. We would like to address this issue in future work.

3.5 Simulations

In our simulation studies, we evaluated the performances of EEScreen and iEEScreen with two different estimating equations, one for a t-year survival model and the other for an accelerated failure time model. We implemented iEEScreen by using EEBoost, as described in Section 3.4, with $\epsilon = 0.01$. We compared these to the naive approach of fitting p_n marginal regressions, as well as to the method of Zhu et al. (2011) and our EEScreen-derived method (3.19) from Section 3.3.

We studied $p_n = 20000$ covariates and set the true parameter vector $\boldsymbol{\beta}_0$ to be such that $\beta_{0j} = 1.5, j = 1, \ldots, 10$, $\beta_{0j} = -0.8, j = 11, \ldots, 20$, and $\beta_{0j} = 0, j = 21, \ldots, p_n$. We generated covariates $\mathbf{X}_i$ from a p_n-dimensional zero-mean multivariate normal. To simulate an easy setting we used a covariance matrix that satisfied the partial orthogonality condition of Fan and Song (2010), where the important covariates were independent of the unimportant covariates. The covariance matrix consisted of 9 blocks of 10 covariates, 1 block of 910

covariates, and 19 blocks of 1000 covariates. Each block had a compound symmetry structure with the same correlation parameter ρ, which was equal to either 0.5 or 0.9, and the blocks were independent from each other. We matched the non-zero components of $\boldsymbol{\beta}_0$ with two of the 10-dimensional blocks. To simulate a more difficult setting we let the entire covariance matrix have a compound symmetry structure with ρ equal to either 0.3 or 0.5.

3.5.1 The t-year survival model

We first considered a t-year survival model. Let T_i and $\mathbf{X}_i$ be the survival time and the covariate vector of the i^{th} patient, respectively. We modeled the probability of surviving beyond some time t_0 conditional on covariates as

$$\text{logit}\{\text{P}(T_i \geq t_0 \mid \mathbf{X}_i)\} = \mathbf{X}_i^T \boldsymbol{\beta}_0. \tag{3.20}$$

This model is very useful in clinical investigations, and in fact we apply it to data from clinical trials of multiple myeloma therapies in Section 3.6.

However, we cannot use the logistic regression because the T_i are not directly observed. Let C_i be the censoring time, such that we only observe $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. Without modeling the C_i, it is difficult to specify a full likelihood model for this data, so we instead turn to estimating equations. To account for the censored data, Jung (1996) assumed that the C_i were independent of the T_i and the $\mathbf{X}_i$ and proposed using the estimating equation

$$\mathbf{U}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \frac{\mathbf{X}_i \pi'(\mathbf{X}_i^T \boldsymbol{\beta})}{\pi(\mathbf{X}_i^T \boldsymbol{\beta})\{1 - \pi(\mathbf{X}_i^T \boldsymbol{\beta})\}} \left\{ \frac{I(Y_i \geq t_0)}{\hat{S}_C(t_0)} - \pi(\mathbf{X}_i^T \boldsymbol{\beta}) \right\}, \tag{3.21}$$

where $\pi(\eta) = \text{logit}^{-1}(\eta)$, $\pi'(\eta) = \partial\pi/\partial\eta$, and $\hat{S}_C(t)$ is the Kaplan-Meier estimate of the survival function of the C_i. According to our procedure, after some simplification we see that EEScreen will retain the parameters

$$\left[j : \left| \sum_{i=1}^{n} X_{ij} \frac{I(Y_i \geq t_0)}{\hat{S}_C(t_0)} \right| \geq \gamma_n \right] \tag{3.22}$$

Table 3.1: Median minimum model size (interquartile range) for the t-year survival model

	Partial orthogonality		Compound symmetry	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.3$	$\rho = 0.5$
EEScreen	2849 (6180)	22 (249.5)	19666.5 (610.5)	19676 (559.5)
Marginal	2908 (6278)	22 (228.75)	19659 (611.5)	19696 (550.5)
Zhu et al. (2011)	9614.5 (9497.75)	2043.5 (7687)	19647.5 (655.5)	19531.5 (737)
Method (3.19)	7559.5 (11737.75)	944.5 (4121.25)	19614.5 (716.75)	19545.5 (726.5)

Though U_j does not satisfy Assumption 13 because of the $\hat{S}_C(t)$ term, Jung (1996) showed that it can be written in the appropriate form, plus a negligible $o_P(1)$ term. To fit the p_n regressions for the marginal screening method we used a simple Newton-Raphson procedure to solve U_j.

Tuning EEBoost and iEEScreen was difficult because commonly used criteria such as AIC or BIC are not defined in the absence of a likelihood. We instead chose to minimize the GCV-type criterion $\widehat{BS}/(1 - n^{-1}\|\hat{\boldsymbol{\beta}}\|_0)^2$, where $\|\hat{\boldsymbol{\beta}}\|_0$ is the number of nonzero components of $\hat{\boldsymbol{\beta}}$, and $\widehat{BS}$ is the estimate of the Brier score at t_0. If $\hat{\pi}(t_0 \mid \mathbf{X}_i)$ is the predicted survival probability of patient i at t_0, then $\widehat{BS}$ is defined by Graf et al. (1999) as

$$\widehat{BS} = n^{-1} \sum_i \left[\frac{\{0 - \hat{\pi}(t_0 \mid \mathbf{X}_i)\}^2}{\hat{S}_C(X_i)} I(Y_i \leq t_0, \delta_i = 1) + \frac{\{1 - \hat{\pi}(t_0 \mid \mathbf{X}_i)\}^2}{\hat{S}_C(t_0)} I(Y_i \geq t_0) \right]. \quad (3.23)$$

We generated survival times for $n = 100$ subjects from $\log(T_i) = \mathbf{X}_i^T \boldsymbol{\beta}_0 + \varepsilon_i$ with ε_i having a logistic distribution with mean -0.5 and scale 1. Under this scheme the model of Jung (1996) is correctly specified. We generated C_i from an exponential distribution to give approximately 50% censoring. We observed that the 20^{th} percentile of the simulated survival times was roughly $t_0 = 0.005$, so we used this t_0 when implementing the estimating equation. We simulated 200 such datasets.

Table 3.1 reports the median sizes of the smallest models $\hat{\mathcal{M}}$ found by the different screening methods that still contained the true model $\mathcal{M}$. The performance is best under the partial orthogonality setting when $\rho = 0.9$, which is not surprising because this setting leads

Table 3.2: Average runtime in seconds (standard deviation) for the t-year survival model

	Partial orthogonality		Compound symmetry	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.3$	$\rho = 0.5$
EEScreen	1.29 (0.09)	1.38 (0.47)	1.38 (0.36)	1.32 (0.16)
Marginal	617.79 (61.99)	1023.79 (1405.58)	1608.09 (2594.27)	1054.86 (252.32)
Zhu et al. (2011)	1.52 (0.08)	1.58 (0.45)	1.88 (4.47)	1.49 (0.2)
Method (3.19)	1.54 (0.09)	1.58 (0.45)	2.13 (8.02)	1.48 (0.18)

to the greatest separation between the important and unimportant covariates. EEScreen and marginal screening show similar performances, while our method (3.19) appears to actually outperform the method of Zhu et al. (2011) in the partial orthogonality setting.

Though EEScreen and marginal screening produce similar results, Table 3.2 shows that marginal screening, at least for this t-year survival model, can take much longer. These simulations were run on the Orchestra cluster supported by the Harvard Medical School Research Information Technology Group, on machines with 3.6 GHz Intel Xeon processors with at least 12GB of memory, and marginal screening took at least 10 minutes. On the other hand, the EEScreen-type methods and the method of Zhu et al. (2011) were completed in a few seconds, showing the EEScreen can be much more computationally efficient than standard screening methods.

To better understand the performances of these various screening methods, we studied in Figure 3.1 the average number of false positives corresponding to a given number of false negatives achieved by the screened model $\hat{\mathcal{M}}$. We again see that the methods perform best in the partial orthogonality setting when the correlation is high. Furthermore, given the same setting, EEScreen performs better than the model-free methods. This is most likely because the model used by EEScreen is correctly specified, and thus should be more powerful than the model-free methods. This type of phenomenon was also pointed out by Zhu et al. (2011). As in Table 3.1, our method (3.19) again appears to outperform that of Zhu et al. (2011).

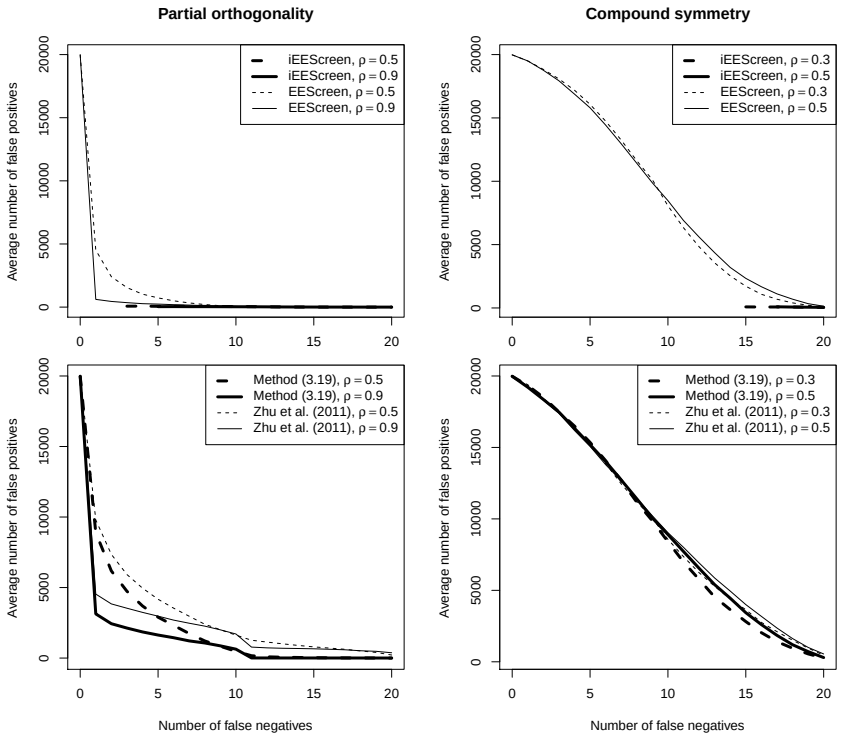

Figure 3.1: Screening performances for the t-year survival model

Figure 3.1 also shows that in all cases, the variable selection performance of iEEScreen far outperforms the other methods, particularly in the compound symmetry setting. However, we found that iEEScreen is not able to include all of the important covariates. In the partial orthogonality setting, it can only include up to 17 or 18 of the important covariates, while in the compound symmetry setting it cannot achieve fewer than 15 false negatives. It turns out that the boosting procedure we use to implement iEEScreen saturates at some point in its fitting, perhaps due to the fact that there are more parameters than covariates, or perhaps because our choice for the boosting parameter $\epsilon = 0.01$ might be too large.

Next we studied the effect on estimation accuracy of using screening before fitting a regularized regression model with EEBoost. Figure 3.2 reports the average mean squared error of estimation (MSE) as a function of $|\hat{\mathcal{M}}|$, the number of variables kept after screening. Here we defined MSE as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2$, where $\hat{\boldsymbol{\beta}}$ is the estimate obtained by EEBoost after screening. It is clear that using EEScreen first can improve the estimation accuracy of

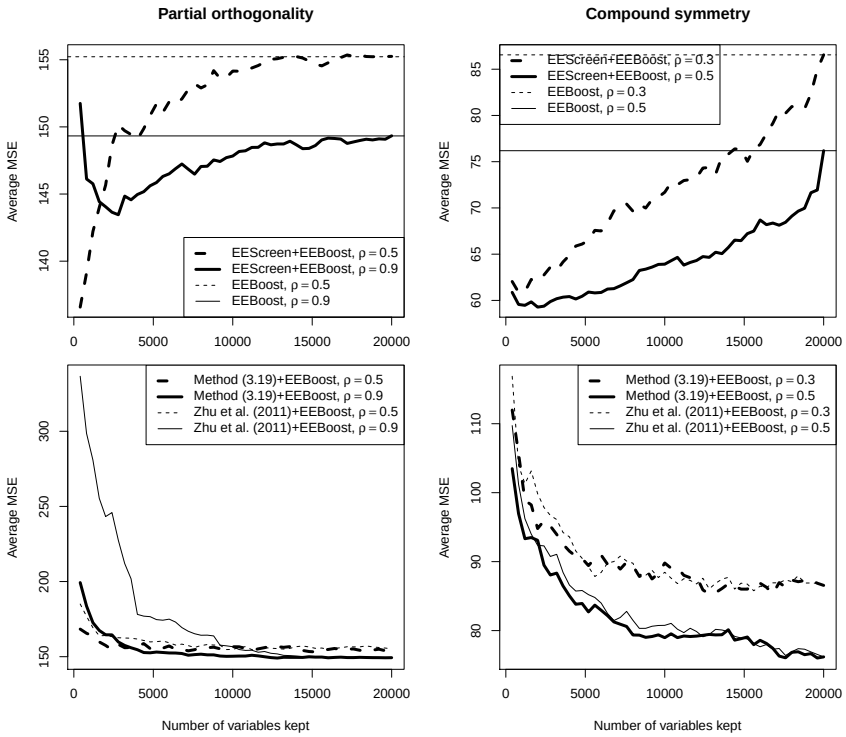

Figure 3.2: Mean squared errors for the t-year survival model

EEBoost, especially in the compound symmetry setting. Screening with the model-free methods does not appear to reduce the MSE, perhaps because they need to retain a large number of covariates before they include the important variables (Table 3.1).

On the other hand, estimation error is not so meaningful in the absence of a correctly specified model. We therefore considered the out-of-sample predictive ability, as measured by the AUC statistic (Uno et al., 2007) at time t_0, of the models fit by EEBoost after screening in Figure 3.3. In the partial orthogonality settings, using EEScreen first does not appear to have much of an effect on the AUC, while in the compound symmetry setting it does improve the predictive ability of the subsequent fitted model. Our model-free method (3.19) does not seem to have much of an effect on AUC in either setting, but appears to perform slightly better than the method of Zhu et al. (2011).

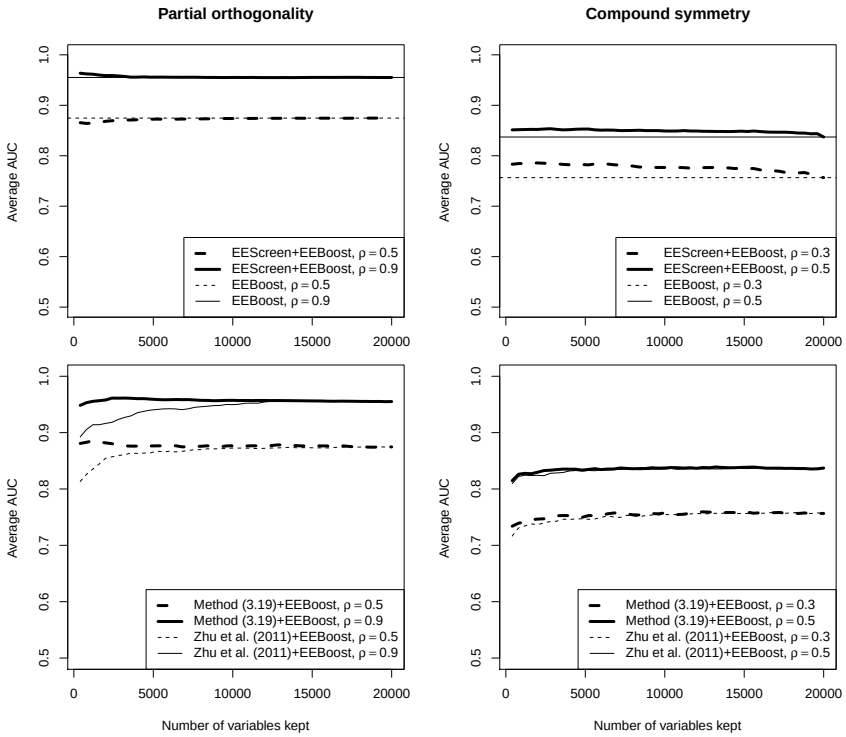

Figure 3.3: Out-of-sample AUCs for the t-year survival model

3.5.2 The accelerated failure time model

The t-year survival model is useful when we are interested in a fixed event time. To study the entire survival distribution, one useful approach is the accelerated failure time (AFT) model, which posits that

$$\log(T_i) = \mathbf{X}_i^T \boldsymbol{\beta}_0 + \varepsilon_i, \tag{3.24}$$

where the ε_i are independent and identically distributed, and the ε_i can have an arbitrary distribution. The $\boldsymbol{\beta}$ can be estimated using the U-statistic-based estimating equation

$$\mathbf{U}(\boldsymbol{\beta}) = n^{-2} \sum_{i=1}^{n} \sum_{k=1}^{n} (\mathbf{X}_k - \mathbf{X}_i) I\{e_i(\boldsymbol{\beta}) \leq e_k(\boldsymbol{\beta})\} \delta_i, \tag{3.25}$$

where $e_i(\boldsymbol{\beta}) = \log(Y_i) - \mathbf{X}_i^T \boldsymbol{\beta}$ (Tsiatis, 1996; Jin et al., 2003; Cai et al., 2009). Following our procedure, after some simplification we see that EEScreen will retain the parameters

$$\left\{ j : \left| \sum_{i=1}^{n} \sum_{k=1}^{n} (X_{kj} - X_{ij}) I(Y_i \leq Y_k) \delta_i \right| \geq \gamma_n \right\}. \tag{3.26}$$

Table 3.3: Median minimum model size (interquartile range) for the AFT model

	Partial orthogonality		Compound symmetry	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.3$	$\rho = 0.5$
EEScreen	997 (2968.75)	20 (2)	19829.5 (316.25)	19822.5 (401.25)
Marginal	1750.5 (3742.25)	21 (144)	19835 (353)	19764 (436)
Zhu et al. (2011)	10761.5 (9416)	747 (3804.5)	19482 (854)	19464.5 (922.5)
Method (3.19)	7940.5 (11962.75)	282.5 (2230.5)	19501.5 (800.25)	19522 (785.75)

This is a U-statistic of order $m = 2$ and therefore satisfies our assumptions in Section 3.2.3. Despite being a discrete estimating equation, (3.25) poses no additional problems to EEScreen or iEEScreen. To fit the p_n regressions for the marginal screening method we used the method of Jin et al. (2003), available in the R package `lss`.

To tune EEBoost and iEEScreen, consider the function

$$L(\boldsymbol{\beta}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \{e_j(\boldsymbol{\beta}) - e_i(\boldsymbol{\beta})\} I\{e_i(\boldsymbol{\beta}) \leq e_j(\boldsymbol{\beta})\} \delta_i. \tag{3.27}$$

Cai et al. (2009), in their work on regularized estimation for the AFT model, argued that $L(\boldsymbol{\beta})$ is an adequate measure of the accuracy of estimation. They and Jin et al. (2003) also noted that $\mathbf{U}(\boldsymbol{\beta})$ is the "quasiderivative" of $-L(\boldsymbol{\beta})$. For these reasons, we tuned EEBoost by minimizing the GCV-type criterion

$$L(\hat{\boldsymbol{\beta}})/(1 - n^{-1}\|\hat{\boldsymbol{\beta}}\|_0)^2, \tag{3.28}$$

where we used $L(\boldsymbol{\beta})$ in place of a negative log-likelihood.

We generated $n = 100$ survival times from $\log(T_i) = \mathbf{X}_i^T \boldsymbol{\beta}_0 + \varepsilon_i$ with ε_i having a standard normal distribution. We generated C_i independently from an exponential distribution to give approximately 50% censoring, and we simulated 200 datasets.

We report for the different screening methods the smallest $\hat{\mathcal{M}}$ that still contained $\mathcal{M}$ in Table 3.3. As with the t-year survival model, the methods perform best in the partial orthogonality setting with $\rho = 0.9$. We also again see that our method (3.19) outperforms

Table 3.4: Average runtime in seconds (standard deviation) for the AFT model

	Partial orthogonality		Compound symmetry	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.3$	$\rho = 0.5$
EEScreen	1.58 (0.15)	1.53 (0.1)	1.51 (0.1)	1.51 (0.11)
Marginal	1024.71 (114.2)	971.85 (82.5)	1081.56 (149.64)	1203.19 (106.81)
Zhu et al. (2011)	1.6 (0.16)	1.46 (0.11)	1.44 (0.1)	1.46 (0.11)
Method (3.19)	1.6 (0.15)	1.46 (0.11)	1.44 (0.09)	1.45 (0.11)

the method of Zhu et al. (2011). In addition, Table 3.4 shows that marginal screening is much more time-consuming than the EEScreen-based methods or the procedure of Zhu et al. (2011).

Figure 3.4 reports the average number of false positives contained in $\hat{\mathcal{M}}$ as the number of allowed false negatives is varied. As in the t-year survival model simulations, iEEScreen performs better than non-iterative EEScreen, though in the compound symmetry case it also saturates before it can select all of the important covariates. We also see that the EEScreen outperforms the model-free methods again, and that our method (3.19) somewhat outperforms the method of Zhu et al. (2011). The plots in Figure 3.4 for the model-free methods look very similar to the corresponding ones in Figure 3.1, and this is because the models used to generate both survival times were both AFT models, differing only in the distributions of the error terms.

The average mean square errors of the models fit after screening are plotted in Figure 3.5. Similar to the results for the t-year survival model, we see that screening using model-free methods does not improve the estimation accuracy of the subsequent regularized regression fit. Interestingly, for the AFT model it appears that screening with EEScreen only barely decreases the MSE under partial orthogonality, and is actually detrimental to the MSE in the compound symmetry setting, in contrast to the results for the t-year survival model.

We see something similar when we examine the out-of-sample predictive abilities of the models fit by EEBoost after screening. We calculated the C-statistics (Uno et al., 2011a)

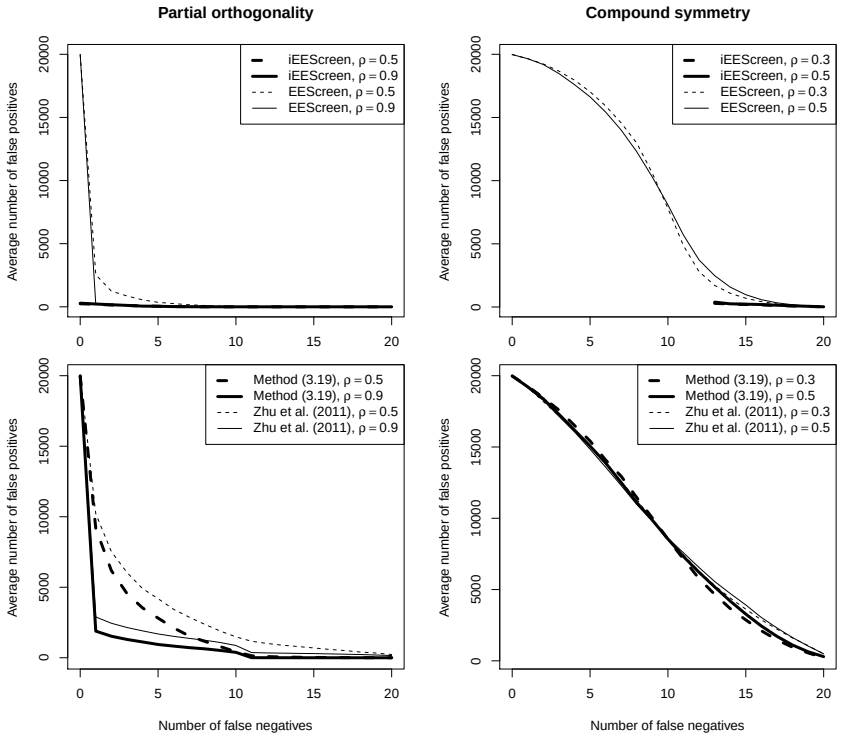

Figure 3.4: Screening performances for the AFT model

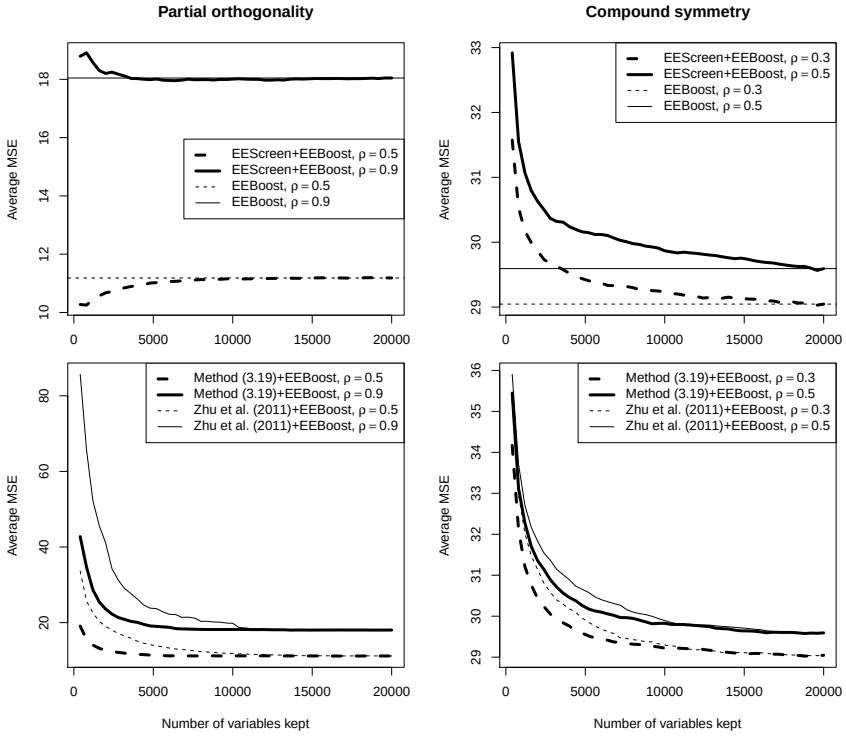

Figure 3.5: Mean squared errors for the AFT model

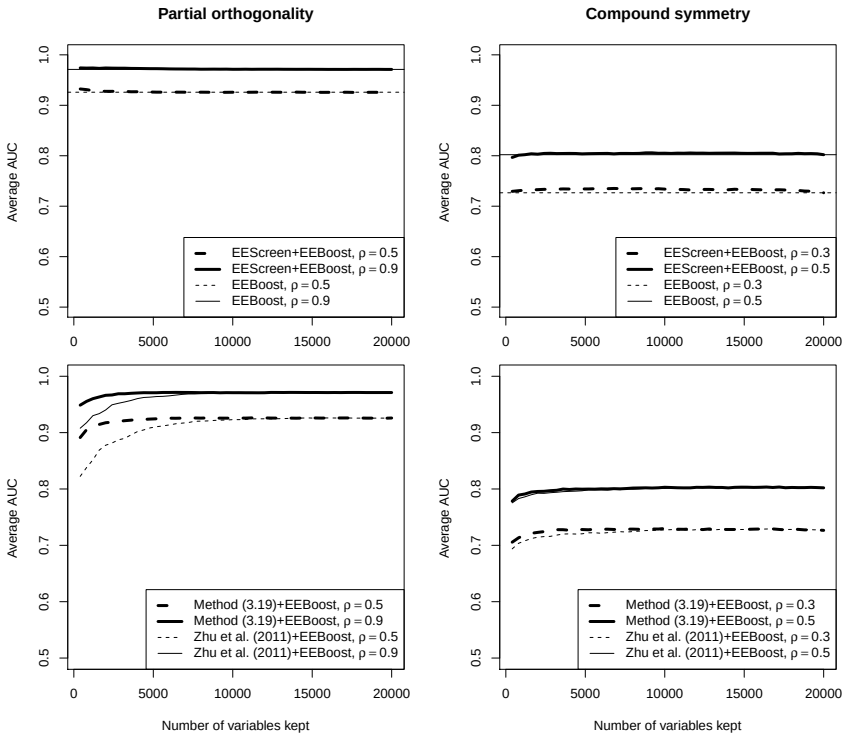

Figure 3.6: Out-of-sample C-statistics for the AFT model

of the fitted models on independently generated datasets and report them in Figure 3.6. EEScreen does not have much of an effect on the C-statistic, while using the model-free methods tend to decrease the predictive ability of the fitted model.

The results in Figures 3.5 and 3.6 are in contrast to the corresponding t-year survival simulation results, which showed the EEScreen can indeed improve MSE and prediction. This may be due to the way these figures were generated: to plot these figures we varied the size of $\hat{\mathcal{M}}$ from between 400 to 20000 in increments of 400. However, the advantages of screening in the AFT setting perhaps may only be seen if fewer than 400 covariates are retained.

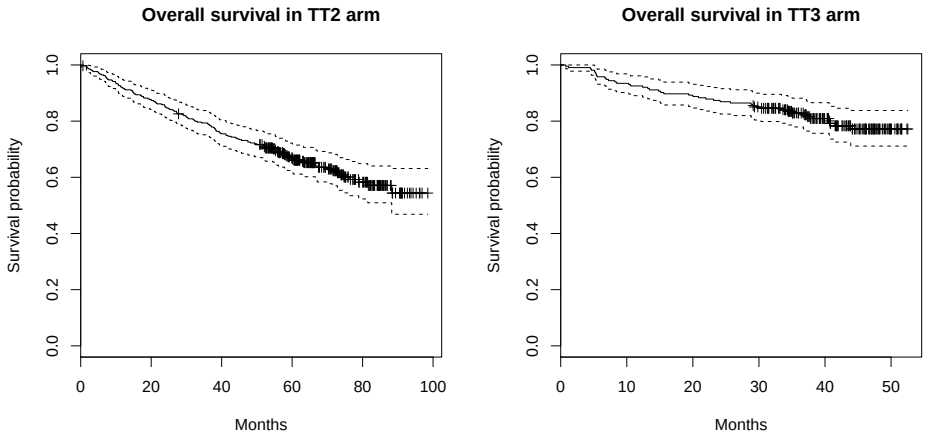

Figure 3.7: Kaplan-Meier estimates from multiple myeloma clinical trials

3.6 Data example

We illustrate our methods on data from a multiple myeloma clinical trial. Multiple myeloma is the second-most common hematological cancer, but despite recent advances in therapy the sickest patients have seen little improvement in their prognoses. It is of great interest to explore whether genomic data can be used to predict which patients will fall into this high-risk subgroup, so that they might be targeted for more aggressive or experimental therapies.

The MicroArray Quality Control Consortium II (MAQC-II) study posed exactly this question to 36 teams of analysts representing academic, government, and industrial institutions (Shi et al., 2010). It used data from newly diagnosed multiple myeloma patients who were recruited into clinical trials UARK 98-026 and UARK 2003-33, which studied the treatment regimes total therapy II (TT2) and total therapy III (TT3), respectively (Zhan et al., 2006; Shaughnessy et al., 2007). Teams were asked to predict the probability of surviving past $t_0 = 24$ months, which is roughly the median survival time of high-risk myeloma patients (Kyle and Rajkumar, 2008), using the TT2 arm as the training set and the TT3 arm as the testing set.

Table 3.5: AUCs for probability of surviving past $t_0 = 24$ months

Method	Optimal $\lvert\hat{\mathcal{M}}\rvert$	5-fold CV AUC (SD)	AUC in TT3
EEScreen (t-year)	5000	0.61 (0.03)	0.61
Method (3.19)	10	0.63 (0.06)	0.58
Zhu et al. (2011)	100	0.67 (0.08)	0.59
EEScreen (AFT)	100	0.65 (0.08)	0.70

There were 340 patients in TT2, with 126 events and an average follow-up time of 55.82 months, and 214 patients in TT3, with 43 events and an average follow-up of 37.03 months. The Kaplan-Meier estimates of the survival curves are given in Figure 3.7. Gene expression values for 54675 probesets were measured for each subject using Affymetrix U133Plus2.0 microarrays, and 13 clinical variables were also recorded, including age, gender, race, and serum β_2-microglobulin and albumin levels.

Figure 3.7 shows that there was a patient in TT2 censored before 24 months, so we cannot model these data using simple logistic regression. We therefore considered the t-year survival model with estimating equation (3.21), from Section 3.5.1. Because we had a total of 54688 covariates and only 340 patients in TT2, we first implemented a screening step, where we considered EEScreen, our model-free method (3.19), and the method of Zhu et al. (2011). We then fit the screened variables using EEBoost, with the generalized cross-validation criterion described in Section 3.5.1. To choose the size of $\hat{\mathcal{M}}$, we used 5-fold cross-validation and selected the value of $|\hat{\mathcal{M}}|$ that gave the best average AUC statistic. The values we considered were 10, 50, 100, 500, 1000, and the numbers from 5000 to 54688 in increments of 5000. Finally, we validated our model in the TT3 arm.

Table 3.5 summarizes our results. We first focused on the AUCs estimated using five-fold cross-validation. Surprisingly, we found that EEScreen gave us the lowest AUC, and that the model-free methods required fewer covariates while giving better prediction. However, note that screening using the t-year survival estimating equation (3.21) essentially dichotomizes the observed times to binary outcomes, because we are only modeling whether they are

larger than t_0. In contrast, we can see from the forms of method (3.19) and the procedure of Zhu et al. (2011) that they use continuous outcomes. We therefore hypothesized that the model-free methods had more power than EEScreen based on equation (3.21) to detect covariate effects, even though they did not incorporate any modeling assumptions.

To test this hypothesis we examined the performance of using EEScreen based on the AFT model estimating equation (3.25). This strategy does not dichotomize the survival outcomes and is also a more restrictive model than the t-year model because it makes a global assumption on the distribution of the survival times. After screening we still used the t-year survival model to fit the retained covariates. Indeed, Table 3.5 shows that with this strategy, we needed to retain only 100 covariates to achieve a high AUC.

Turning now to the validation AUCs calculated in the TT3 arm, we found that though the model-free methods gave higher AUCs in cross-validation, their validation AUCs were essentially comparable to that of EEScreen based on the t-year survival model. This might perhaps indicate that the model-free methods actually overfit to patients in the TT2 arm, and thus their results didn't generalize well to patients treated with TT3. In contrast, the EEScreen method based on the AFT model gave a much higher validation AUC of 70%. The final fitted model contained 37 covariates, which in addition to various gene expression levels also included β_2-microglobulin, albumin, and lactate dehydrogenase levels. Thus our method was able to select important clinical predictors in addition to identifying potentially important genomic factors.

3.7 Discussion

In this paper we introduced EEScreen, a new computationally convenient screening method that can be used with any estimating equation-based regression method. We proved finite-sample performance guarantees that hold for any model that can be fit with U-statistic-based

estimating equations, and in addition showed that our approach could be used to derive a model-free screening procedure very similar to one proposed by Zhu et al. (2011). Finally, we have drawn a connection between screening and boosting methods, showing that the EEBoost algorithm of Wolfson (2011) can be viewed as a form of iterative screening.

Our simulation results, conducted using a t-year survival model as well as the AFT model, support the use of EEScreen in practice. They suggest that EEScreen is capable of retaining most of the important covariates without also including too many false positives, unless the covariates are very highly correlated. In terms of estimation and prediction, when the working model is correctly specified, using EEScreen will usually not give worse results than not using screening at all, and at the very least will dramatically reduce the required computation time. This does not always appear to be true of the model-free methods.

On the other hand, in our multiple myeloma example we saw that using different models for the screening step and the regression step can offer better performance than keeping to one model throughout. This illustrates the difficulty in choosing a default screening procedure that works well in all cases. However, our myeloma results suggest that one key consideration is the power of the screening step. The AFT model-based screening appeared to have greater power than the t-year model, and perhaps its modeling assumptions prevented it from overfitting to the TT2 arm, as the model-free methods seemed to do.

This insight implies that different situations will require choosing different screening methods in order to achieve the greatest power. Estimating equations give us access to a wide range of models to choose from, with more parametric models offering lower variance but higher bias, and models with fewer assumptions offering the opposite tradeoff. Thus our EEScreen approach is perfectly suited to this screening strategy, offering quick computation and good theoretical properties for whichever model we decide to use.

Acknowledgments

We thank Professors Lee Dicker and Julian Wolfson for reading an earlier version of this manuscript. We also thank Professors Tianxi Cai, Tony Cai, Jianqing Fan, Hongzhe Li, and Xihong Lin for their many helpful comments and suggestions.

3.8 Appendix A: Proofs

3.8.1 Proof of Theorem 7

The event $\{\mathcal{M} \subseteq \hat{\mathcal{M}}\}$ equals $\{\min_{j\in\mathcal{M}} |U_j(0)| \geq \gamma_n\}$, so it is easy to see that

$$\mathrm{P}(\mathcal{M} \subseteq \hat{\mathcal{M}}) \geq 1 - \sum_{j\in\mathcal{M}} \mathrm{P}(|U_j(0)| < \gamma_n). \tag{3.29}$$

By the triangle inequality, we know that for all j, $|u_j(0)| \leq |U_j(0) - u_j(0)| + |U_j(0)|$, and by Assumption 15 we see that $c_1[n/m]^{-\kappa} - |U_j(0)| \leq |U_j(0) - u_j(0)|$ for all $j \in \mathcal{M}$. Therefore, $|U_j(0)| < \gamma_n$ for $j \in \mathcal{M}$ implies $|U_j(0) - u_j(0)| \geq c_1[n/m]^{1-\kappa}/2$. We can conclude from Assumptions 13 and 14 and Bernstein's inequality for U-statistics (Hoeffding, 1963) that

$$\mathrm{P}(\mathcal{M} \subseteq \hat{\mathcal{M}}) \geq 1 - 2s_n \exp\left\{-\frac{c_1^2[n/m]^{1-2\kappa}/4}{2\Sigma^2 + bc_1[n/m]^{-\kappa}/3}\right\} \tag{3.30}$$

3.8.2 Proof of Theorem 8

For the marginal estimating equations U_j and their expected values u_j, we know from Assumptions 13 and 14 and Bernstein's inequality for U-statistics (Hoeffding, 1963) that

$$\mathrm{P}\{\max_j |U_j(0) - u_j(0)| \leq c_1[n/m]^{-\kappa}/4\} \geq 1 - 2p_n \exp\left\{-\frac{c_1^2[n/m]^{1-2\kappa}/16}{2\Sigma^2 + bc_1[n/m]^{-\kappa}/6}\right\}. \tag{3.31}$$

Also, if $\max_j |U_j(0) - u_j(0)| \leq c_1[n/m]^{-\kappa}/4$, then $|U_j(0)| \geq \gamma_n$ implies that $|u_j(0)| \geq c_1[n/m]^{-\kappa}/4$. This means that

$$|\hat{\mathcal{M}}| = |\{j : |U_j(0)| \geq \gamma_n\}| \leq |\{j : |u_j(0)| \geq c_1[n/m]^{-\kappa}/4\}| \leq \frac{16}{c_1^2[n/m]^{-2\kappa}} \sum_j u_j(\mathbf{0})^2. \quad (3.32)$$

From our EEScreen procedure described in Section 3.2.1, we see that the $u_j(\mathbf{0})$ are the possibly relabeled components of the expected full estimating equation $\mathbf{u}(\mathbf{0})$. Thus $\sum_j u_j(\mathbf{0})^2 = \|\mathbf{u}(\mathbf{0})\|_2^2$, and by the generalization of the mean value theorem to vector-valued functions (Hall and Newell, 1979) and Assumptions 17 and 16,

$$\|\mathbf{u}(\mathbf{0})\|_2 = \|\mathbf{u}(\boldsymbol{\beta}_0) - \mathbf{u}(\mathbf{0})\|_2 \leq \sup_{0<t<1} \|\mathbf{i}(t\boldsymbol{\beta}_0)\|_2 \|\boldsymbol{\beta}_0\|_2 \leq c_2 \sup_{0<t<1} \sigma_{\max}\{\mathbf{i}(t\boldsymbol{\beta}_0)\} = c_2\sigma^*_{\max}, \quad (3.33)$$

so that

$$\mathrm{P}\left[|\hat{\mathcal{M}}| \leq \frac{16c_2^2\sigma^{*2}_{\max}}{c_1^2[n/m]^{-2\kappa}}\right] \geq \mathrm{P}\{\max_j \|U_j(0) - u_j(0)\|_\infty \leq c_1[n/m]^{-\kappa}/4\} \quad (3.34)$$

$$\geq 1 - 2p_n \exp\left\{-\frac{c_1^2[n/m]^{1-2\kappa}/16}{2\Sigma^2 + bc_1[n/m]^{-\kappa}/6}\right\}. \quad (3.35)$$

References

Adelstein, D. J., Li, Y., Adams, G. L., Wagner, H., Kish, J. A., Ensley, J. F., Schuller, D. E., and Forastiere, A. A. (2003). An intergroup phase III comparison of standard radiation therapy and two schedules of concurrent chemoradiotherapy in patients with unresectable squamous cell head and neck cancer. *Journal of Clinical Oncology* **21,** 92–98.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.-C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., and Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology* **24,** 537–544.

Antoniadis, A., Fryzlewicz, P., and Letué (2010). The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics* page to appear.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Ser. B* **57,** 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29,** 1165–1188.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24,** 2350–2383.

Bunea, F., Wegkamp, M., and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference* **136,** 4349–4364.

Cai, J., Fan, J., Li, R., and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92,** 303–316.

Cai, T., Huang, J., and Tian, L. (2009). Regularized estimation for the accelerated failure time model. *Biometrics* **65,** 394–404.

Cai, T., Wang, L., and Xu, G. (2010). Shifting inequality and recovery of sparse signals. *IEEE Transactions on Signal Processing* **58,** 1300–1308.

Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics* **35,** 2313–2351.

Conn, A., Scheinberg, K., and Vicente, L. (2009). *Introduction to derivative-free optimization*, volume 8. Society for Industrial Mathematics.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Ser. B* **34,** 187–220.

Decaux, O., Lodé, L., Magrangeas, F., Charbonnel, C., Gouraud, W., Jézéquel, P., Attal, M., Harousseau, J. L., Moreau, P., Bataille, R., Campion, L., Avet-Loiseau, H., and Minvielle, S. (2008). Prediction of survival in multiple myeloma based on gene expression profiles reveals cell cycle and chromosome instability signatures in high-risk patients and hyperdiploid signatures in low-risk patients: a study of the Intergroupe Francophone du Myélome. *Journal of Clinical Oncology* **26,** 4798–4805.

Dharmadhikari, S. W., Fabian, V., and Jogdeo, K. (1968). Bounds on the moments of martingales. *The Annals of Mathematical Statistics* **39,** 1719–1723.

Dicker, L. (2011). *Regularized regression methods for variable selection and estimation.* PhD thesis, Harvard University, Boston, MA.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* **106,** 544–557.

Fan, J., Feng, Y., and Wu, Y. (2010). Ultrahigh dimensional variable selection for Cox's proportional hazards model. *IMS Collections* page to appear.

Fan, J. and Li, R. (2001). Variable selection via noncave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348–1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* **30,** 74–99.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Ser. B* **70,** 849–911.

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research* **10,** 2013–2038.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models and NP-dimensionality. *The Annals of Statistics* page to appear.

Fine, J. P., Ying, Z., and Wei, L. J. (1998). On the linear transformation model for censored data. *Biometrika* **85,** 980–986.

Fleming, T. R. and Harrington, D. P. (2005). *Counting processes and survival analysis*. Wiley, Hoboken.

Friedman, J., Hastie, T., Hoefling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **2,** 302–332.

Fu, W. (2003). Penalized estimating equations. *Biometrics* **59,** 126–132.

Gorst-Rasmussen, A. and Scheike, T. H. (2011). Independent screening for single-index hazard rate moels with ultra-high dimensional features. Technical Report R-2011-06, Department of Mathematical Sciences, Aalborg University.

Graf, E. A., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18,** 2529–2545.

Gu, M. (1992). On the Edgeworth expansion and bootstrap approximation for the cox regression model under random censorship. *Canadian Journal of Statistics* **20,** 399–414.

Hadzidimitriou, A., Stamatopoulos, K., Belessi, C., Lalayianni, C., Stavroyianni, N., Smilevska, T., Hatzi, K., Laoutaris, N., Anagnostopoulos, A., Kollia, P., and Fassas, A. (2006). Immunoglobulin genes in multiple myeloma: expressed and non-expressed repertoires, heavy and light chain pairings and somatic mutation patterns in a series of 101 cases. *Haematologica* **91,** 781–787.

Hall, W. S. and Newell, M. L. (1979). The mean value theorem for vector valued functions: a simple proof. *Mathematics Magazine* **52,** 157–158.

Hideshima, T., Mitsiades, C., Tonon, G., Richardson, P., and Anderson, K. C. (2007). Understanding multiple myeloma pathogenesis in the bone marrow to identify new therapeutic targets. *Nature Reviews Cancer* **7,** 585–598.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58,** 13–30.

Hofmann, W.-K., de Vos, S., Komor, M., Hoelzer, D., Wachsman, W., and Koeffler, H. P. (2002). Chracterization of gene expression of $CD34^+$ cells from normal and myelodysplastic bone marrow. *Blood* **100,** 3553–3560.

Huang, A., Xu, C., and Wang, M. (2011). A modified SLP algorithm and its global convergence. *Journal of Computational and Applied Mathematics* **235,** 4302–4307.

Huang, J., Ma, S., Xie, H., and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96,** 339–355.

Huang, Y.-T., Heist, R. S., Chirieac, L. R., Lin, X., Skaug, V., Zienolddiny, S., Haugen, A., Wu, M. C., Wang, Z., Su, L., Asomaning, K., and Christiani, D. C. (2009). Genome-wide analysis of survival in early-stage nonsmall-cell lung cancer. *Journal of Clinical Oncology* **27,** 2660–2667.

James, G. M. and Radchenko, P. (2009). A generalized Dantzig selector with shrinkage tuning. *Biometrika* **29,** 323–337.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90,** 341–353.

Johnson, B. A., Lin, D. Y., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103,** 672–680.

Johnson, B. A., Long, Q., and Chung, M. (2011). On path restoration for censored outcomes. *Biometrics* (in press).

Jung, S.-H. (1996). Regression analysis for long-term survival rate. *Biometrika* **83,** 227–232.

Kyle, R. and Rajkumar, S. (2008). ASH 50th anniversary review: Multiple myeloma. *Blood* **111,** 2962–2972.

Lagakos, S. (2006). The challenge of subgroup analyses – reporting without distorting. *New England Journal of Medicine* **354,** 1667–1669.

Lesaffre, E. and Marx, B. D. (1993). Collinearity in generalized linear regression. *Communications in Statistics – Theory and Methods* **22,** 1933–1952.

Leyffer, S. and Mahajan, A. (2010). Nonlinear constrained optimization: methods and software. Technical Report ANL/MCS-P1729-0310, Argonne National Laboratory.

Li, G., Peng, H., Zhang, J., and Zhu, L. (2011). Robust sure independence screening for ultrahigh dimensional models. Technical Report arXiv:1012.4255v2.

Li, H. (2008). Censored data regression in high-dimensional and low-sample-size settings for genomic applications. In Biswas, A., Datta, S., Fine, J., and Segal, M., editors, *Statistical Advances in Biomedical Sciences: State of the Art and Future Directions*, pages 384–403. Wiley, Hoboken.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* **84,** 1074–1078.

Liu, H., Zhang, J., Jiang, X., and Liu, J. (2009). The group dantzig selector. In *International Conference on Artificial Intelligence and Statistics*.

Mackinnon, M. J. and Puterman, M. L. (1989). Collinearity in generalized linear models. *Communications in Statistics – Theory and Methods* **18,** 3463–3472.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Statistics* **28,** 863–884.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis* **52,** 374–393.

Mulligan, G., Mitsiades, C., Bryant, B., Zhan, F., Chng, W. J., Roels, S., Koenig, E., Fergus, A., Huang, Y., Richardson, P., Trepicchio, W. L., Broyl, A., Sonneveld, P., Shaughnessy, J. D., Bergsagel, P. L., Schenkein, D., Esseltine, D. L., and Boral, A. (2007). Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. *Blood* **109,** 3177–3188.

Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems, NIPS-2009*.

Oken, M. M., Creech, R. H., Tormey, D. C., Horton, J., Davis, T. E., McFadden, E. T., and Carbone, P. P. (1982). Toxicity and response criteria of the Eastern Cooperative Oncology Group. *American Journal of Clinical Oncology* **5,** 649–655.

Sarkar, S. K. (2004). FDR-controlling stepwise procedures and their false negatives rates. *Journal of Statistical Planning and Inference* **125,** 119–137.

Shaughnessy, J., Zhan, F., Burington, B. E., Huang, Y., Colla, S., Hanamura, I., Stewart, J. P., Kordsmeier, B., Randolph, C., Williams, D. R., et al. (2007). A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood* **109,** 2276–2284.

Shaughnessy, J. D. and Barlogie, B. (2003). Interpreting the molecular biology and clinical behavior of multiple myeloma in the context of global gene expression profiling. *Immunological Reviews* **194,** 140–163.

Shi, L., Campbell, G., Jones, W. D., et al. (2010). The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology* **28,** 827–838.

Spiekerman, C. F. and Lin, D. Y. (1998). Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association* **93,** 1164–1175.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73,** 363–369.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Ser. B* **58,** 267–288.

Tibshirani, R. J. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16,** 385–395.

Tibshirani, R. J. (2009). Univariate shrinkage in the Cox model for high dimensional data. *Statistical Applications in Genetics and Molecular Biology* **8,** 21.

Tsiatis, A. A. (1996). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics* **18,** 354–372.

Ueki, M. (2009). A note on automatic variable selection using smooth-threshold estimating equations. *Biometrika* **96,** 1005–1011.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011a). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30,** 1105–1117.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011b). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30,** 1105–1117.

Uno, H., Cai, T., Tian, L., and Wei, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* **102,** 527–537.

van de Geer, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics* **23,** 1779–1801.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer, New York.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory* **55,** 2183–2202.

Wang, L., Zhou, J., and Qu, A. (2011). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* .

Wang, S., Nan, B., Zhou, N., and Zhu, J. (2009). Hierarchically penalized Cox regression with grouped variables. *Biometrika* **96,** 307–322.

Wang, S., Nan, B., Zhu, J., and Beer, D. (2008). Doubly penalized buckley–james method for survival data with high-dimensional covariates. *Biometrics* **64,** 132–140.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *The Annals of Statistics* **37,** 2178–2201.

Wolfson, J. (2011). EEBoost: a general method for prediction and variable selection based on estimating equations. *Journal of the American Statistical Association* **106,** 296–305.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Ser. B* **68,** 49–67.

Zhan, F., Huang, Y., Colla, S., Stewart, J., Hanamura, I., Gupta, S., Epstein, J., Yaccoby, S., Sawyer, J., Burington, B., et al. (2006). The molecular classification of multiple myeloma. *Blood* **108,** 2020.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika* **94,** 691–703.

Zhang, H. H., Lu, W., and Wang, H. (2010). On sparse estimation for semiparametric linear transformation models. *Journal of Multivariate Analysis* **101,** 1594–1606.

Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis* **105,** 397–411.

Zhou, B., Peyton, M., He, B., Liu, C., Girard, L., Caudler, E., Lo, Y., Baribaud, F., Mikami, I., Reguart, N., et al. (2006). Targeting ADAM-mediated ligand cleavage to inhibit HER3 and EGFR pathways in non-small cell lung cancer. *Cancer Cell* **10,** 39–50.

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106,** 1464–1475.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.

Zou, H. and Hastie, T. (2005). Regression shrinkage and selection via the elastic net with application to microarrays. *Journal of the Royal Statistical Society, Ser. B* **67,** 301–320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *The Annals of Statistics* **36,** 1509–1533.